DOCUMENT RESUME

ED 212 650	TM 820 024
AUTHOR	Quellmalz, Edys; And Others
TITLE	Studies In Test Design: Annual Report.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY	National Inst. of Education (ED), Washington, D.C.
PUB DATE	Nov 81
GRANT	NIE-G-80-0112
NOTE	323p.; For related documents, see ED 211 592 and TM 820 026.
EDRS PRICE	MF01/PC13 Plus Postage.
DESCRIPTORS	Cost Effectiveness; Criterion Referenced Tests; Elementary Secondary Education; Higher Education; Learning Processes; *Measurement Techniques; Pictorial Stimuli; Research Utilization; Responses; Scoring; Student Placement; *Test Construction; *Testing Problems; *Test Validity; *Writing Evaluation; Writing Instruction; Writing Skills
IDENTIFIERS	Inter Rater Reliability

ABSTRACT

This document contains the following manuscripts: "Effects of Alternate Scoring Options on the Classification of Entering Freshmen Writing Competencies," by Edys Quellmalz and Eva Baker; "Implications of Learning Research for Designing Competency Based Assessment," by Edys Quellmalz; "Effects of Alternative Discourse and Response Modes on Characterizations of Students' Writing Performance," by Frank Capell, Edys Quellmalz and Chi Ping Chou; "Problems in Stablizing the Judgment Process," by Edys Quellmalz; "Effects of Visual or Written Topic Information on Essay Quality," by Eva Baker and Edys Quellmalz; "Effects of Time and Strategy Use on Writing Performance," by Linda Polin; "Designing Writing Assessments: Balancing Fairness, Utility and Cost," by Edys Quellmalz; "The Measurement of Students' Writing Performance in Relation to Instructional History," by Marcella Pitts; "Measures of High School Students' Expository Writing: Direct and Indirect Strategies," by Laura Spooner-Smith; and "Alternative Scoring Systems for Predicting Criterion Group Membership," by Lynn Winters. (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Center for the Study of Evaluation

UCLA Graduate School of Education
Los Angeles, California 90024

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED212650

TM820 024

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. C. Bear

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Deliverable - November 1981

STUDIES IN TEST DESIGN

Annual Report
Edys Quellmalz, Project Director

Grant Number
NIE-G-80-0112
P-3

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

This document contains manuscripts prepared by the Test Design Project for refereed publications.

Test Design: Studies in Writing Assessment

Annual Report
November 1981

Effects of Alternate Scoring Options on the Classification of Entering Freshmen Writing Competencies

- Edys Quellmalz and Eva Baker

Implications of Learning Research for Designing Competency Based Assessment

- Edys Quellmalz

Effects of Alternative Discourse and Response Modes on Characterizations of Students' Writing Performance

- Frank Capell, Edys Quellmalz and Chi Ping Chou

Problems in Stablizing the Judgment Process

- Edys Quellmalz

Effects of Visual or Written Topic Information on Essay Quality

- Eva Baker and Edys Quellmalz

Effects of Time and Strategy Use on Writing Performance

- Linda Polin

Designing Writing Assessments: Balancing Fairness, Utility and Cost

- Edys Quellmalz

The Measurement of Students' Writing Performance in Relation to Instructional History

- Marcella Pitts

Measures of High School Students' Expository Writing: Direct and Indirect Strategies

- Laura Spooner-Smith

Alternative Scoring Systems for Predicting Criterion Group Membership

- Lynn Winters

EFFECTS OF ALTERNATIVE SCORING OPTIONS ON THE CLASSIFICATION OF ENTERING FRESHMEN WRITING COMPETENCIES

Edys Quellmalz and Eva Baker

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Among the many criticisms of the quality of public education, complaints about students' inability to write prose lead the pack. At the time of college admission when students need to be assigned to beginning English courses, writing deficiencies become especially salient. At entrance to college, students may be assigned to college-level beginning English courses, or with greater frequency, may be placed in a special course designed to remedy composition problems and to prepare for regular college level work. This initial placement decision is made through different means. Some schools base their decision solely on student verbal scores on a college entrance examination. Others require that all students take a special placement examination. These examinations may vary in their development history (locally prepared or commercially published), definition of writing (narrative or expository prose), format (multiple choice or essay production), and manner by which the passing score is determined. An ideal and experimentally clean way to make choices among such alternatives would involve the systematic variation of some of these variables to determine which procedures provide the least mistaken estimate of students' writing ability. In fact, admission is a serious business and little experimental "fooling" with the system is tolerated in real colleges and universities, even for the promised benefit of improved decisions.

This study, however, is an attempt to contrast alternative assessment methods in actual placement testing. Its practical impetus grew from specific requirements in the higher education system in California. As background, California has two, state-wide university systems: The University of California (UC) and the California State University and Colleges (CSUC). Although the systems are designed to attract different levels of students (at UC, the top 12½% statewide and at CSUC, the top 33%) students may transfer from system to system or to different campuses within the same system. CSUC consists of 19 campuses, and to standardize requirements among campuses, a committee of faculty cooperated with the Educational Testing Service (ETS) to develop a system-wide test of English composition placement, the English Placement Test (EPT). The UC system of nine campuses operates so that each campus' unique placement test (called the Subject A examination) is honored by the other campuses. Since CSUC students often wish to transfer to UC schools, a study group made up of faculty from both systems was appointed to review the need for common writing placement procedure for all UC and CSUC campuses. The use of the English Placement Test was suggested by the CSUC representatives.

The problem in its most simple form is whether the EPT would provide the same quality of information thought to be obtained through the existing procedures at UC campuses. Could a test designed for a population consisting of the top one-third of students operate efficiently for the top 12½%?

Embedded in this problem are a number of serious issues related to the teaching and testing of writing. For a start, few agree on the definition of writing competence itself. A common, but operationally vague desire is that students ought to write well enough to succeed in other college courses, as if success were an unidimensional phenomenon. In fact, Smith (1975) demonstrated that requirements for success vary from college specialization to specialization. Definitions of competence may focus on particular features of writing, such as structural or grammatical elements. In other views, acceptable mechanics are a minimum, but emphasis is given, in addition, to the quality of thought or to the logic and clarity of the communication.

A second issue running through this study is the form of student response used to make the decision. Some tests of writing rely heavily on "indirect" measurement, where performance on multiple choice tests is used to "predict" writing achievement. These tests are justified along these connected lines of argument. First, the correlation coefficients of written essays and multiple choice tests are high enough that the "validity" of the objective test should not be challenged. The tests are functionally thought to measure the "same thing" (Godshalk, Swineford, & Coffman, 1966; Breland & Braucher, 1977). Given this equivalence, efficiency favors choosing the least expensive method, and objective tests are easier and cheaper to administer and score. The scoring argument is bolstered by the well-known differences in raters' judgments of essays, that is, the matter of scorer unreliability.

Proponents of collecting writing samples from students argue that the cognitive requirements of creating essays and answering a series of multiple choice tests differ markedly from one another, and that no amount of statistical modelling can actually equate writing with choosing the right answer (Spooner-Smith, 1978; Quellmalz & Capell, 1979). Further criticisms of rater unreliability are countered by the results of good training procedures. However, the cost issue remains, cast by these advocates as a choice between cheap, irrelevant information or more costly, valid data.

A third issue applies to any definition or format for the assessment of student writing competence: how are standards of passing or failing set? Does the standard treat equally the two forms of potential misclassification, competent students who "fail" and incompetent students who "pass"? Is there a policy that the benefit of the doubt goes to the student? Does the system so value its definition of writing that it wishes to be conservative about who gets to enter college English courses?

A last, but critical issue arises for those who have opted for the collection of essay responses. Not only questioned are the number, type, and length of responses necessary for accurate judgment, but also heated disagreement occurs over the best scoring procedures. The choices are between holistic scoring, which gives an overall estimate of the essay, and analytic scoring which provides subscores for particular characteristics of the writing. Again, the conflict is between cost, where holistic scoring takes approximately 2/3 the time of analytic scoring, and precision of information, where analytic scores provide diagnosis of deficient performance. Strong advocates for holistic scoring cite its economy (Godshalk, et al., 1966; Alloway, 1978; Powills, Bowers & Conlan, 1979). However, feature analyses of good and poor papers point to the distinct differences in their content and structure (See Cooper, Cherry, Gerber, Fleischer, Copley, & Sartisky, 1979), and advocates of analytic ratings argue for the use of such information in determining instructional policy for remediation (Quellmalz, 1980).

With contention as a backdrop, then, the practical problem of choosing a "good" placement procedure for UC was studied. Staff at a university-based research center proposed research to compare three alternative methods for making the placement decision: the use of the English Placement Test (EPT) (consisting of an essay and multiple choice scales) proposed by the CSUC staff; the placement procedure (Subject A examinations) in use at each of the two UC campuses; an analytic essay rating scale developed by the research center in the course of its studies of writing (the CSE scale). Two simple questions were formulated to guide this study:

1. How comparable are the scores students receive from each form of writing assessment?
2. Would the methods sort students in competent and incompetent groups in the same way?

METHODS

Overview

Each of two UC campuses agreed to participate in the study. Instead of requiring their own Subject A examination, each campus administered the EPT examination to a sample of students participating in regular placement examinations. The EPT essay was first scored by ETS, rescored at each campus using campus scoring procedures (both campuses used holistic rating procedures), and then the essays were sent to the research center for rerating according to the CSE analytic scheme. Actual placement decisions for each student were made on the basis of the campus interpretation of ETS scores.

Subjects

Three hundred eight high school seniors were required to take the experimental version of the placement examination at either of two UC campuses. A placement test for writing was a regular requirement for students scoring between 450 and 600 on the College Entrance Examination Board (CEEB) test.

Instruments

The English Placement Test

The EPT was developed by the Educational Testing Service in collaboration with CSUC as a placement tool for first-year English classes in the CSUC system. The EPT requires students to write one 45-minute essay and to complete a 90-minute multiple choice section covering three skill areas: reading, sentence construction, and logic and organization. The reading section asks students to identify main ideas and to interpret ideas in short reading passages. The sentence construction test items require students to recognize arrangements of sentence elements that "express meaning clearly and correctly." The logic and organization section contains a variety of item types intended to measure students' ability to "see relationships between words." For example, some items require students to arrange words into categories; other items involve identifying sentences to begin, end, or support a given paragraph. Still other items intend to measure the students' ability to distinguish between fact and opinion. The objective part of the EPT counts 75% of the total.

Essay topic. The essay direction required students to write a 45-minute essay on a topic eliciting narrative/descriptive writing. The topic of this administration called for students to write about "a real or an apparent change that had occurred in someone they knew."

EPT essay criteria. The EPT scoring scale is a six-point holistic essay scale divided into two parts--"upper half papers" and "lower half papers." Raters are instructed to read each paper through quickly and assign an overall rating based on how well the essay addressed itself to all aspects of the question (topic), how well the essay is organized, and how well it demonstrates writing quality. Aspects of writing quality mentioned in the rubric are syntax and diction. Papers that do not respond to, argue or avoid the question are scored zero. The EPT was studied for content validity, as reported by Breland and Ragosa (1976). Unfortunately, no results were available.

UC Campus 1: holistic essay criteria

Campus 1 employed a six-point holistic scale which permits readers to assign a plus or minus to each point on the scale (1=high, 6=low). The rubric directs raters' attention to the thesis statement and its development, sentence structure, word choice, and a detailed list of "mechanics" features. Additionally, each point on the scale corresponds to a placement decision. For example, scores of one, two or three indicate that the student is prepared to take a regular freshman composition course, while a score of four through six indicates that the student should be placed in one of a series of increasingly remedial English classes. Campus 1 typically employs a one-hour placement examination.

Campus 2: holistic essay criteria

A six-point holistic rating scale was also employed by Campus 2 (1=low, 6=high). The rubric emphasizes fluency and mechanics, although reference is made to the logic and organization of the writing scale. In its normal placement examination, two one-hour essays are produced by each student at Campus 2.

CSE analytic essay criteria

Unlike the three holistic approaches of the other rating procedures, the CSE essay scoring provides an analytic rating of each essay (Quellmalz, 1979). The analytic rubric derived from other scales used for narrative discourse and from texts and tests in composition and rhetoric (Pitts, 1978). The scale presents carefully explicated criteria developed for domain-referenced narrative writing tasks. Scale criteria require reference to observable features in an essay, unlike many rating rubrics which include more subjective, affective judgments. The scale consists of five subscales, each with a range of four points. Based on studies suggesting that holistic and analytic ratings provide distinct information about student writing, the scale calls for both holistic and analytic ratings (Winters, 1978). The first subscale, General Impression, directs raters to read the paper quickly first and to rate it according to their global judgments of its quality as an example of narration. The remaining four subscales attend to the following components of the writing: focus, organization, support, and mechanics. The scoring rubric for the scale contains a detailed description of essay features associated with each of the four levels of quality within each of the subscales.

Archival student information

In addition to the three scores generated by the rescoring of the required placement exam, Scholastic Aptitude Test (SAT) verbal scores, College Entrance Examination Board (CEEB) scores, High School English course grades and grade point averages were also available for students.

Procedures

Administration

Students who came to the required UC placement examination were divided, as they arrived, into groups taking the regular or the experimental EPT administration. Students in the study were placed in the same room and not exposed to the usual campus procedure. The entire EPT was administered according to the publisher's directions. This process was repeated on each of the two UC campuses in the study.

EPT Scoring Procedure

The essays rated by the EPT procedures were graded at the same time as a larger pool of essays from all CSUC campuses (n=6,293). Twenty-seven raters were trained in a three and one-half hour training session to assign scores according to the EPT rubric. Each essay was read by two readers and the final score assigned to an essay was the sum of the two scores. As the EPT rubric was a six-point scale, essay scores ranged from one to twelve. Papers with scores differing by two or more points and all papers that received a zero score from one reader but a non-zero score from the other reader were read by a third reader. The total essay score in these adjudicated cases was the sum of the two most congruent scores. EPT reported that the majority of discrepant scores occurred in the three to five score range.

Rater agreement was calculated by a correlation coefficient summarizing the amount of agreement between the first and second scores assigned to a paper, rather than of the amount of agreement between particular rater pairs. The correlation coefficient reported for 5,756 papers was .59.

CSE Rating Procedures

The combined set of 308 essays was rescored at the research center using the CSE Factual Narrative Scale II. Four raters, English instructors, were hired to read the essays. All of the raters had previous experience in the systematic rating of student essays, and two of the four raters had used the particular scale in previous studies.

CSE rater training procedures were similar to those employed by Spooner-Smith (1978) and Winters (1978). Approximately four hours were devoted to review, rating and discussion of 30 sample essays on the essay topic. At the conclusion of the training session, rater agreement coefficients were computed for each of the subscores and the total scale in order to determine whether training should be continued. Alphas ranged from .86 to .92 (based on four ratings per paper), and generalizability coefficients ranged from .59 to .87. As a result, readers reread and discussed the pilot test papers again for the one subscale with low reliability, focus, before reading the actual "experimental" essays. Papers were randomly assigned to raters.

Campus 1: rating procedure

Six teaching assistants experienced in teaching basic writing rated the Campus 1 essays returned by ETS. The Campus 1 scale, based primarily on a tally of mechanical errors, was used to assign essay scores. Each paper was read by one reader; raters were department teaching assistants and were given no additional formal training.

Campus 2: rating procedure

Campus 2 papers were read by seven raters, all composition instructors. The raters had previous experience in rating placement essays for the English department, so only about one and a half hours were devoted to rater training. During this session, raters read and discussed essays on topics analogous to the EPT topic and assigned scores according to the Campus 2 writing exam scale.

Each paper was read by two raters; the final score was the sum of the two ratings. Papers discrepant by two or more points were read by a third reader and the discrepancy resolved in the same manner as were discrepancies in the EPT scoring procedure. Campus 2 calculated no interrater reliabilities.

RESULTS

Comparability of Assessment Procedures

The first section of results addresses the comparability of the three alternative measures and includes internal analyses of each (see Table 1). The EPT and CSE scores will be treated first because they each provide subscales. Consider the EPT analyses. The most dramatic findings surround the relationship of the objective EPT subscales and the essay score (see Table 2). Each of three subscales strongly correlates with one another, a fact which suggests that they may provide redundant information. These subscales, taken individually or combined into an "objective" composite relate only moderately with the EPT essay score analyses (ranges of r between .25 and .30).

- - - - - - - - - - - - - - - - -

Insert Tables 1 & 2 about here

- - - - - - - - - - - - - - - - -

The CSE scale analysis addresses the relationship of the four analytic subscores, the total of these scores, and the General Impression, "holistic" score for each essay (see Table 3). The relatively low correlations suggest that the particular subscales are, in fact, identifying separate skill

- - - - - - - - - - - - - - -

Insert Table 3 about here

- - - - - - - - - - - - - - -

Table 1

Means and Standard Deviations

	Possible	n	$\overline{X}$	s.d.	n	$\overline{X}$	s.d.
EPT TOTAL	180	104	152.38	6.44	201	154.17	3.84
EPT ESSAY	12	104	7.03	1.38	201	7.37	1.58
EPT OBJECTIVE SCALES							
Reading	180	104	153.04	8.81	201	154.20	12.10
Sentence construction	180	104	154.72	7.35	201	156.01	11.89
Logic and organization	180	104	153.16	7.89	201	154.01	11.95
Composition	180	104	152.03	6.13	201	153.09	11.53
Total Objective Score	540	104	460.92	22.11	201	464.21	35.00
CAMPUS SCORING		103	2.93	1.26	201	6.61	2.02
CSE SUBSCALES							
General Impression	4	69	1.52	.71	148	1.79	.78
Focus	4	69	1.80	.56	148	1.98	.55
Organization	4	69	1.70	.63	148	2.00	.70
Support	4	69	1.88	.67	148	2.09	.69
Mechanics	4	69	1.91	.53	148	2.35	.62
Total	20	69	8.81	2.35	148	10.17	2.52

TABLE 2

Internal Characteristics of EPT and CSE Assessment

English Placement Test

EPT	Essay	Reading	Sentence construction	Logic	Composition	Objective	Total
Essay							
Reading	.27						
Sentence construction	.28	.68					
Logic	.25	.71	.62				
Composition	.71	.70	.79	.78			
Objective Total	.30	.91	.86	.88	.85		
Total	.62	.85	.81	.81	.97	.93	

N= 308

TABLE 3

Center for the Study of Evaluation analytic scale

CSE Scale	General Impression	Focus	Organization	Support	Mechanics	Total
General Impression						
Focus	.47					
Organization	.75	.47				
Support	.48	.46	.55			
Mechanics	.46	.41	.32	.28		
Total	.85	.72	.83	.73	.65	

N = 217

components. The correlation of .85 for the General Impression and the total of the subscales suggests that directing one's attention to four particular features of writing nonetheless produces values consistent with an overall holistic view.*

The comparison between features assessed by the EPT and the CSE indicators more directly addresses the question of assessment comparability (see Table 4).

- - - - - - - - - - - - - -
Insert Table 4 about here
- - - - - - - - - - - - - -

The essay scores derived from EPT and CSE scoring suggest that only a moderate amount of overlap exists in the scoring rubrics. The holistic ratings between the CSE General Impression and EPT essay correlate in the mid-ranges; however, the component skills measured by the CSE analytic dimensions and the EPT subscales diverge dramatically. For instance, "organization" is assessed by both EPT and CSE scores, yet the correlation between subscales is only .12. Sentence construction on the EPT and mechanics on the CSE subscale, apparently comparable dimensions, correlate .29. Clearly, the format of the EPT subscale responses (objective tests) assesses a different capacity than the CSE subscale rating of the essay.

Comparisons were also made among the EPT scores, CSE scores, and the UC campus holistic scoring procedures. In Table 5, the first column pre-

- - - - - - - - - - - - - -
Insert Table 5 about here
- - - - - - - - - - - - - -

---

*In fact, the holistic score is undoubtedly contaminated by the raters' use of the analytic rating scales, after the first paper, that is.

TABLE 4

Cross-Correlations Between EPT and CSE Subscales

Campuses Combined

EPT	CSE General Impression	Focus	Organization	Support	Mechanics	Total
Essay	.46	.46	.41	.42	.38	.56
Reading	.17	.15	.14	.16	.27	.23
Sentence construction	.18	.18	.16	.15	.29	.25
Logic & organization	.14	.20	.12	.11	.23	.21
Composition	.36	.39	.32	.31	.40	.4
Objective test	.19	.20	.16	.16	.30	.27
Total	.39	.33	.28	.28	.39	.42

TABLE 5

Correlation of Placement Test Scores from EPT, Campus 1 Campus 2, and CSE

	EPT essay	EPT objective	Campus 1	Campus 2	CSE
EPT essay					
EPT objective	.30				
Campus 1	.60	.53			
Campus 2	.25	.08	*		
CSE	.40	.27	.48	.12	

*Campus 1 and 2 scored only their own students' essays.

18

sents the simplest contrasts. The EPT correlates at the .40 level with the CSE total. The holistic scoring procedures at the UC campuses results in discrepant relationships (at Campus 1, r=.60, and at Campus 2, r=.25). A low risk conclusion is that "holistic" ratings (as used at each campus and for the EPT rating) mean different things. In any case, inferences about the stability of these relationships is certainly weakened by the relatively low inter-rater reliability reported for the EPT ratings, the lack of reliability estimates for the UC efforts, and the potential for error inherent in the single rating procedure used at Campus 1. Yet, even if these ratings were reliable, the conclusion from these data would be that raters using different systems operationalize writing in very different ways.

Relationship of assessment procedure and archival information

Table 6 presents descriptive statistics for archival data by campus and Table 7 displays the correlations among different writing assessment methods and other writing-related archival data often used in placement decisions. Making inferences from such spotty results is dangerous; however, the most consistent relationships are among the College Entrance, Scholastic Aptitude, and English Placement Tests. While this relationship may result from connections between underlying abilities (for instance, comprehension ability is assessed on all three measures), one might argue that the fact that these tests originate from the same publisher, using

- - - - - - - - - - - - - - - - - -

Insert Tables 6 & 7 about here

- - - - - - - - - - - - - - - - - -

TABLE 6

Means and Standard Deviations
for Archival Data* by Campus

	Campus 1			Campus 2		
	N	$\overline{X}$	s.d.	N	$\overline{X}$	s.d.
High School English Grades	61	3.68	.34	187	3.70	.35
High School Grade Point Average	90	3.68	.28	161	3.68	.28
College Entrance Examination Board	90	478	79	184	509	63
Scholastic Aptitude Test (Verbal)	90	492	87	180	510	83

TABLE 7

Correlations Between Alternative Placement Scores and Other Predictors of College English Performance

	College Entrance Examination Board	Scholastic Aptitude Test (Verbal)	High School English	High School Grade Point Average
EPT total score				
Campus 1	.54	.59	.14	.20
Campus 2	.66	.64	.32	.31
Combined	.62	.62	.19	.25
CSE total essay score				
Campus 1	.26	.25	-.04	-.01
Campus 2	.29	-.01	.05	.23
Combined	.32	.21	.00	.07
Campus essay score				
Campus 1	.22	.23	.07	.01
Campus 2	.50	.31	.20	.31

supposedly similar test development technology, may be as plausible a link among them.

More disheartening, however, is the lack of relationship among writing indices and high school and English grade point average. Although range restriction definitely must be considered (all students have a 3.2 minimum grade point average to qualify for UC admission), one would still hope that the grades of these students, drawn as they were from the middle of the CEEB distribution (450-600 scores), might support the validity of the measures. One gloomy view is that high school performance, as measured by grades, does not include much writing competence. Research on the amount of actual precollegiate writing required of students supports this analysis (Pitts, 1978).

A related question is the amount of performance that can be inferred to be a specific skill and the amount inferred to be general ability or perhaps general information. The relatively higher values for the Campus 2 procedures may be explained as general ability. This explanation is especially interesting in the light of the weak categories in the scoring rubric, and the form of rater training. When no need exists for identification and operational statement of criteria in order to achieve set levels of agreement among raters, it is reasonable to infer that the writers' general ability rather than specific writing skill is detected by the rating.

Alternative placement decisions using three assessment models

To compare the utility of the three methods in view of different standards for pass and fail, two analyses were performed: 1) the pass score was set at the mean of the scores from the experimental UC distribution; 2) the cut score set according to present or recommended practice. The best approach for identifying the optimal placement of such standards would naturally depend upon developing an adequate estimate of "future success" in college writing, and working back from it, to identify the minimum requirements for competency. In the absence of such a refined external criterion, the alternative placement analyses shed light on the differences in decisions made by the various assessment approaches.

Group analyses

At the group level of analysis, Table 8 displays percentages of students who would be placed in remedial classes if cut-off scores were 1) set at the mean of the UC sample for each of the three methods or 2) set at the recommended or regularly used standard. When the cut-off for the EPT essay is set at the UC mean (a customary ETS procedure), 54% of UC students would be required to take remedial English. If the EPT cut-off score were set at the average of the CSUC population, only 26% of the UC sample would be placed into remedial English. This contrast reflects the differences in populations in the two university systems and suggests that if the EPT essay (and its cut-off) were adopted directly from CSUC, then the standard of writing expected at UC would drop. The CSE scale would place 61% of UC students in the remedial course, with either the average or a substantively set criterion score of 10.

- - - - - - - - - - - - - - -

Insert Table 8 about here

- - - - - - - - - - - - - - -

TABLE 8

Percent of Students Placed in Subject A by the Three Scoring Systems

	When cut-off scores = UC mean			When cut-off scores = those previously used		
Combined campuses	N	Score	Remedial English	N	Score	Remedial English
EPT essay	304	< 7.28	54	304	≤ 6	26
EPT total	304	<153.62	48	304	≤150	18
CSE total	235	< 9.83	61	235	≤ 10	61
Campus 1						
Campus rubric	103	< 2.93	49	104	≤ 4	31
EPT essay	103	< 7.03	63	104	≤ 6	34
EPT total	103	<152.38	35	104	≤150	20
CSE total	71	< 8.61	51	71	≤ 10	79
Campus 2						
Campus rubric	201	< 6.61	40	200	≤ 7	40
EPT essay	201	< 7.37	50	200	≤ 6	23
EPT total	201	<154.27	43	200	≤150	14
CSE total	164	< 10.35	53	164	≤ 10	53

Contrasts in performance between the two UC campuses demonstrate that Campus 2 apparently draws from a somewhat more proficient population of writers than Campus 1.

Individual placement decisions

Different predictions can be made about the placement of any individual student under the three assessment methods (see Table 9). Numbers in the "off" diagonal represent students who would pass under one system and fail according to another (taking pairs of procedures one at a time for each campus). For example, at Campus 1, if the pass score were set at the CSUC mean, 30% of the students who pass the EPT essay would fail using the regular standards of the campus, and 57% would fail using the CSE scale. Placement discrepancies between CSE and Campus 1 procedures are greater than between Campus 1 and the EPT decisions. Campus 2 placement decisions similarly demonstrate discrepancies, but with different details. For instance, in comparing the CSE with Campus 2 standards, one can see that 36% of the students would pass in one system and fail in the other. However, the degree of difficulty (as judged by the percentages passing and failing in either system) shows rough equivalence. Thus, in the case of the Campus 2-CSE comparison, it is the defintion of writing competency that accounts for differences in placement rather than "difficulty" of the measure.

- - - - - - - - - - - - - -

Insert Table 9 about here

- - - - - - - - - - - - - -

TABLE 9

Comparison of Placements When Essay Cut-off Scores Are Set at Previously Employed Standards

Campus 1 rubric

EPT essay rubric	Pass ≤3	Fail ≥4	
Pass ≥7	37	31 (30%)	68
Fail ≥6	31 (30%)	4	35
	68	35	103

Campus 2 rubric

EPT essay rubric	Pass ≥8	Fail ≤7	
Pass ≥7	79	76 (38%)	155
Fail ≤6	9 (5%)	36	45
	88	112	200

CSE rubric

Campus 1 rubric	Pass ≤11	Fail ≥10	
Pass ≤3	5	40 (57%)	45
Fail ≥4	10 (14%)	15	25
	15	55	70

CSE rubric

Campus 2 rubric	Pass ≥11	Fail ≤10	
Pass ≥8	47	29 (18%)	76
Fail ≤7	30 (18%)	58	88
	77	87	164

CSE rubric

EPT essay rubric	Pass ≥11	Fail ≤10	
Pass ≥7	15	33 (47%)	48
Fail ≤6	0 (0%)	23	23
	15	56	71

CSE rubric

EPT essay rubric	Pass ≥11	Fail ≤10	
Pass ≥7	70	57 (35%)	127
Fail ≤6	7 (4%)	29	36
	77	86	163

DISCUSSION

The findings of the study dramatize the dilemma facing multi-site educational systems attempting to establish uniform writing competency testing. The question is whether newly proposed placement method B is better than extant placement method A, and the answer is, in this case unfortunately, "It depends." It depends on what you are looking for and what evidence will convince you that you have found it. This study underscores the fact that writing is not an undifferentiated skill construct and that different tests may measure or emphasize very different aspects of the writing competency domain.

The questions guiding this study structured information about the consequences of using different assessment methods: 1) Are descriptions of student writing competence provided by the proposed placement exam comparable to campus methods in use or to an analytic essay scoring scheme? and 2) Do alternative placement methods result in the same placement decisions? The answer to both of these questions is, basically, "No."

The data indicate that descriptions of a student's writing competence derived from the three alternative measures, the EPT (essay and objective tests), the local campus rubrics, and the CSE essay scale differ considerably. These differences are indicated by the generally low correlations among the placement methods and other writing-related indices, and, most importantly, by the discrepant classification of the same student as master or non-master. These empirical analyses suggest a need to return to a logical and psychological analysis of the content of the three measurement approaches as they relate to what is meant by writing competence.

The low or moderate correlations of the ratings generated by the EPT, UC campus and CSE rubrics imply that the criteria in these scales emphasize different essay features. A look at the content of the rubrics confirms these differences. Even when nominally similar methods were used, empirical differences were found. For instance, both the EPT and Campus 2 rubrics were applications of the ETS holistic scoring procedures applied in large scale writing assessments (Conlan, 1976; Alloway, 1978; Powills, et al., 1979). Yet the same basic approach results in clearly different specifications and applications of criteria by different sets of raters. These results, at minimum, challenge the stability and validity of holistic scoring for placement and competency decisions, where it is critical that consistent criteria be applied fairly to all students.

Our data illustrate that, contrary to folklore, competent writing does not "surface" apart from the details of the rating scheme. The view of writing competency reflected in any rating procedure vastly influences what happens to students. The results of this study were presaged by earlier work. In a study of the effects of alternative response criteria in holistic, analytic and quantitative rating schemes, Winters (1978) also found that the scales differentially profiled the same set of essays and characterized students as masters or non-masters. Furthermore, she reported that imprecisely worded criteria were refined and clarified by raters during training, and she hypothesized that a new set of raters would refine and apply the criteria differently.

This study suggests that the design of writing placement assessments require detailed and systematic consideration of a range of test development issues. Methodology for designing domain-referenced tests (DRT) in general (Hively, 1974; Baker, 1974; Popham, 1978, 1980) and for domain-referenced writing assessment in particular (Quellmalz, 1978, 1980; Baker & Quellmalz, 1979) may provide a useful approach to developing or selecting writing assessments. Such methods begin with a detailed definition of desired writing competencies and then require precise domain specifications for the rhetorical features of the writing task, explicit criteria in the rating scale, and reliable procedures for using the scale. These specifications permit examinations of the planned placement test by subject matter and testing experts prior to the test administration. For example, screening of the task structure and scoring procedures in this study might have resulted in changing the essay task from a narrative one to an expository task more representative of the type of writing required in college courses. Examination of the planned scoring methods might have resulted in the calculation of interrater reliability for Campus 2 and for the scoring of placement essays by more than one rater for Campus 1.

The design of the domain of task and scoring features for a particular placement test also can provide a blueprint for guiding development of comparable, parallel writing tasks, rating criteria and rating procedures, assuring the fairness of decisions from occasion to occasion and site to site. In the ideal case, evidence should indicate that the placement test discriminates between surviving and floundering college writers. This study emphasizes the need for a systematic approach to selecting or developing writing competency tests. Perhaps through domain-referenced testing methods and continuing longitudinal research on writing assessment problems, we can improve the confidence we place in decisions about writing ability.

References

Alloway, J. E. Some ways of establishing criteria for assessing writing performance from the perspective of the test developer. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

Baker, E. L., & Quellmalz, E. S. *Results of pilot studies*. Report to the National Institute of Education, Los Angeles: UCLA Center for the Study of Evaluation, 1979. (OB-NIE-G-78-0213)

Baker, E. L. Beyond objectives: Domain-referenced tests for evaluation and instructional improvement. *Educational Technology*, 1974, *14*, 10-21.

Breland, H. M., & Braucher, J. L. Measuring writing ability. Paper presented at the annual meeting of the American Educational Research Association, New York, 1977.

Breland, H., & Ragosa, D. Validating placement tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

Conlan, G. How the essay in the CEEB English test is scored. Princeton, N.J.: Educational Testing Service, 1976.

Cooper, C., Cherry, R., Gerber, R., Fleischer, S., Copley, B., & Sartisky, M. Writing abilities of regularly-admitted freshmen at SUNY/Buffalo. University Learning Center and Graduate Program in English Education, Department of Learning and Department of English, State University of New York, Buffalo, 1979.

Godshalk, F. I., Swineford, F., & Coffman, W. E. *The measurement of writing ability*. New York: College Entrance Examination Board, 1966.

Hively, W. Introduction to domain-referenced testing. *Educational Technology*, 1974, *14*, 5-10.

Pitts, M. *The relationship of classroom instructional characteristics and writing in the descriptive/narrative mode*. Report to the National Institute of Education, Los Angeles: UCLA Center for the Study of Evaluation, 1978. (Grant No. OB-NIE-G-78-0213)

Popham, W. J. *Criterion-referenced measurement*. Englewood Cliffs, N.J.: Prentice-Hall, 1978.

Popham, W. J. Domain-referenced strategies. In R. A. Berk (Ed.), *Criterion-referenced measurement*. Johns Hopkins University Press, 1980.

Powills, J. A., Bowers, R., & Conlan, G. Holistic essay scoring: An application of the model for the evaluation of writing ability and the measurement of growth in writing ability over time. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Quellmalz, E. Assessing writing proficiency: Designing integrated multi-level information systems. Paper presented at the annual meeting of the National Reading Conference, San Diego, CA, 1980.

Quellmalz, E. Defining writing domains: Effects of discourse and response mode. Interim report to the National Institute of Education, Los Angeles: UCLA Center for the Study of Evaluation, 1979. (Grant No. OB-NIE-G-78-0213)

Quellmalz, E. Domain-referenced specifications for writing proficiency. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

Quellmalz, E., & Capell, F. Defining writing domains: Effects of discourse and response mode. Report to the National Institute of Education, Los Angeles: UCLA Center for the Study of Evaluation, 1979. (OB-NIE-G-78-0213)

Smith, L. An assessment of writing needs of undergraduates in the life sciences and social sciences divisions at UCLA. Unpublished thesis, University of California, Los Angeles, 1975.

Spooner-Smith, L. Investigation of writing assessment strategies. Report to the National Institute of Education, November, 1978. (Grant No. OB-NIE-6-78-0213 to the UCLA Center for the Study of Evaluation)

Winters, L. The effects of differing response criteria on the assessment of writing competence. Report to the National Institute of Education, Los Angeles: UCLA Center for the Study of Evaluation, 1978. (Grant No. OB-NIE-G-78-0213)

IMPLICATIONS OF LEARNING RESEARCH FOR DESIGNING COMPETENCY BASED ASSESSMENT

Edys Quellmalz

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

IMPLICATIONS OF LEARNING RESEARCH FOR DESIGNING COMPETENCY BASED ASSESSMENT

One might argue that theory and methodology in learning research should generally precede and inform theory and practice in instruction and testing. Although such linear patterns of research and development rarely are found in practice, findings in learning can affect test design. This section explores the relationship of learning research and test design in order to provide insight into current problems and to suggest future directions for test design. The section begins by briefly tracing the roots of domain-referenced testing in behavioral psychology, and how current problems in design may be attributed both to superficial application of behaviorism and to the limits of that learning paradigm. The section then presents implications for test design of more recent cognitive processing learning research.

The reader should note that this review reflects the perspective that testing should be integrated with instruction and should inform instructional decision making. This perspective is derived largely from learning and instruction research paradigms and asserts that both test and instructional tasks should be formed from the same or compatible specifications. Further, to be maximally useful, these specifications should reflect learning research and describe precisely the content and response limits for both testing and instructional tasks. In other words, task variables which affect student performance must be considered and specified in instructional and test design.

Domain-Referenced Testing and Behavioral Psychology

As noted in previous sections, domain referenced testing finds its origins in Skinner's operant theory of human behavior (Skinner, 1954). Skinner's analysis of learning identified two salient elements in learning: a stimulus and an overt behavioral response elicited by the stimulus. Mental processing of stimulus information by the learner was excluded from the learning model since internal, unobservable phenomena were characterized as inaccessible to hypothesis testing. The methodology and theory developed in this hypothesis testing resulted in several interrelated principles applicable to domain-referenced testing and instruction:

1. Stimulus and response requirements must be rigorously described;
2. Learning and criterion tasks must be replicable;
3. Learning tasks must match criterion tasks (or test tasks).

The requirements for carefully specifying task content and response limits advocated by DRT proponents (Hively, 1974; Baker, 1974; Popham, 1980; Polin & Baker, 1979) thus can be derived directly from the experimental paradigm of behaviorism. However, while these researchers call for replicable, rigorous test task specification, practice has generally not heeded the call. Most curriculum programs and commercially published tests, for example, have been developed from vague and imprecise specification and do not attend to content and response dimensions that affect student performance (Anderson, 1972; Wardrop, Anderson, Hively, Anderson, Hastings, & Muller, 1978; Royer & Cunningham, 1978; Quellmalz, 1980). Practice generally has ignored, also, the demand for congruent learning and test tasks. Critiques of instructional programs and curriculum embedded tests document mismatches between instructional objectives and tests purporting to measure those objectives (Baker & Spooner-Smith, 1977; Quellmalz, Snidman, & Herman, 1977). Even learning researchers apparently have not attended very closely to requirements for task and criterion match: Anderson's (1972) review of 130 research articles found that less than 33% gave any rationale for test selection or development, and 51% provided not information about the relationship of content and response requirements for test items and experimental conditions. Montague (1980) similarly noted that reading researchers paid inadequate attention to the alignment of independent and dependent variables. General awareness of the need for precise test and task specification, then, has been lacking.

A second major problem in the state of the art DRT methodology has been inattention to the cognitive processing operations required for students to produce the observable specified response. This problem can be traced to behaviorism's research focus upon overt, observable behaviors and its exclusion of covert, hypothetical processes mediating those responses, although response hierarchies were proposed (Gagné, 1977). The new wave of cognitive psychologists considers the exclusion of cognitive processing in a learning model an error of omission. Critics of objectives-based technology and behaviorsim repeatedly cite the need for methodologies of learning, instruction, and testing to identify, specify, and account for response complexity in terms of information processing which the student must activate to engage the content of the task and produce the requested response (Chomsky, 1957); Greeno, 1976).

The recent shift in learning research from stimulus-response behaviorism to cognitive learning is producing findings with profound implications for the design of compatible test and instructional tasks. Learning has become viewed not as a passive reaction, but as a constructive one. Researchers have resurrected notions from Gestalt psychology and reanalyzed Bartlett's constructionist account of retention (Bartlett, 1932). The learning act is now described as "interactive" (Solomon, 1980) or "generative" (Wittrock, 1974). Research in perception regarding the influence of "anticipatory schema" of what was visually perceived (Nessier, 1967) blends with a schema theory of learning (Anderson, 1977). Language research (Chomsky, 1957) and perceptual research (Nessier, 1967; Paivio, 1971) suggest that learners extracted and abstracted generic rules, strategies, or knowledge representations from repeated encounters with a corpus of oral language, text or problems with recurring elements. These hypothesized internal "frames" (Minsky, 1975) are called schemata, defined as data structures for representing generic, stereotypic concepts stored in memory (Rumelhart & Ortony, 1977). Schemata both determine what information is retained or produced and what it "means." Emphasis has shifted from the end product, a response, to the operations required for the response to occur.

The Cognitive Research Paradigm

Much of learning research in the structure of knowledge in the domains of math, reading, writing, oral language, and artificial intelligence is now attempting to identify the types and sequences of processes, operations, or routines that demark different levels of developing skill in these domains. A salient feature of the paradigm is rigorous description and manipulation of the task content, and intensive analyses of the steps or operations leading to the response. Less often is responding elicited in a recognition, selected response, format than in a production format. Often research methods include extensive protocol analysis of processes reported by the learner.

Types of Knowledge

In an attempt to categorize and form a taxonomy of learning task types, cognitive learning researchers make some distinctions among types of knowledge. These distinctions have important implications for analysis and thus specification of the content and response limits of test tasks. Earlier work (Gagné, 1977) had proposed types of content: knowledge, facts, concepts, principles, and a corresponding response hierarchy. Cognitive researchers propose similar content-response distinctions. Brown, Campione, and Day (1980) write that knowledge divides into three types: strategic, content or factual, and meta-cognitive. Hays-Roth (1980) delineates five types of knowledge: general cognitive skills, declarative knowledge, procedural knowledge, motivation, and attitudes and beliefs. In their discussion of measurement problems in reading comprehension, Royer and Cunningham (1978) distinguish among domain skills, topical world knowledge, and reasoning ability. Other researchers characterize information processing involving two components: a knowledge or content structure component containing a network of concepts and relations, and a set of cognitive processes for operating on the content (Anderson, 1977; Greeno, 1976; Norman & Rumelhart, 1975; Chi & Glaser, 1980).

Common throughout these characterizations of knowledge is the distinction between content and operations, parallel to the distinction made in domain-referenced task specifications. However, these general typologies begin to suggest even more dimensions that could apply to learning, instruction, and test tasks specification. Most critical for the formulation of valid tasks is an emerging awareness of and attempt to separate content and operations integral to mastery of skills in a subject matter domain from content and operations irrelevant to that domain.

Variables in Cognitive Learning Research Relevant to Test Design

With its focus on processes and the conditions influencing them, different variables assumed prominence in cognitive learning research. The remaining parts of this section will discuss the variables most relevant for domain-referenced test design.

Antecedent conditions. Cognitive learning researchers renewed their attention to perceptions, values, social, cultural, and language experiences and to information processing strategies the learner brought to the learning task. In particular, these foci emphasized the critical influence of antecedent conditions on the learners' engagement in the immediate task (Flavell, 1977). Researchers asked, what features of the task did the learner attent to; what schemata, scripts, and plans did the learner have as he engaged in a new task? These cognitive processing questions expanded behaviorists' concern with "entry level skills," instructional specialists' concern with pretesting and, by implication, test designers' attention to item difficulty. Some researchers suggested that the influence of antecedent conditions implied a need for "tailored tests" (Rudner, 1978) or tests sensitive to individual differences in "world knowledge" and in previously established content and response schema. Other research attempted to account or provide for antecedent conditions through more refined design of the learning task.

Because of the diverse task dimensions that cognitive research suggests are important for eliciting performance, the research relating to task elicitation conditions will be discussed as it pertains to separate dimensions within the content limits and response limits of tasks.

Content Limits

1) Context. Among the most striking findings of cognitive learning research has been the sensitivity of performance to the "context" in which the task is set. "Context" includes not just the physical setting, but the social setting in which the learning (and testing) takes place. Writing and oral language researchers criticize the decontextualized features of most school language tasks (instructional and assessment) (Olson, 1977; Scribner & Cole, 1978; Cazden, 1974; Britton, Burgess, Martin, McLeod, & Rosen, 1975; Florio, 1979). They cite the importance of perceived relevance and communicative intent in "real" speech acts and writing acts as important factors in determining learner motivation and deployment of particular strategies. Knowledge of a real and specific purpose and audience differently influence what content, form, and style speakers and writers mobilize (e.g., Olson, 1977; Britton et al., 1975). To the extent that the occasion is natural, familiar, and meaningful to the learner, the writing or speech is more or less motivated and facilitated. When language performance is elicited in decontextualized settings, described by Britton as typical assessment settings, then other, perhaps, irrevelant variables intrude

into the task (Britton et al., 1975).

In addition to information about audience and function, the context also includes a time allocation for task completion. Ability tests often attempt to differentiate individuals through "power" or speeded testing. Formal achievement assessment often sets arbitrary time limits for task performance. Cognitive research on information processing related to subprocesses in a task (e.g., decoding in reading, sentence construction in writing) suggests that less proficient learners might require more time for subroutines which have become automated routines for masters (Nold, 1979; Norman & Bobrow, 1975; Stallard, 1978). Thus time limitations for assessment tasks may not permit students at varying competency levels to complete processes they can perform, e.g., decode and comprehend; plan, write, edit; analyze and solve a problem. Some researchers (see Cooper, 1979) suggest that formal, time-limited assessment should be replaced by sampling of work completed under more "natural" classroom or life conditions.

2) Topicality. In addition to context, the influence of a second task feature, "topicality" or "world knowledge," has become a popular issue in cognitive learning research. World knowledge refers to networks (or a "memory bank") of information about world phenomena learners have in their repertoire. In the subject domains of reading, writing, oral language, and math, world knowledge is considered a prerequisite vehicle for exercise of comprehension, syntactic, or problem solving skills. As Royer and Cunningham (1978) point out, a student cannot apply a main idea identification strategy to a passage about a topic if the student has insufficient topic knowledge to discriminate between superordinate and subor-

dinate pieces of information. Other research documents that lack of topic familiarity or a "script" may result in reduced comprehension and retention of passage content, e.g., Bransford and Johnson (1972), Anderson, Reynolds, Schallert, and Goetz (1977).

Similarly, in written production, a student cannot begin to compose a coherenet essay without a sufficient store of facts and relations within a topic. Topical content of test passages or of writing topics can be differentially biased against students with particular cultural or language experinces (Capell, 1980; Royer & Cunningham, 1978; Baker & Quellmalz, 1980). Thus topic familiarity is a critical task feature in test design. This dimension may be provided for in test specifications by either empirically verifying subjects' topic familiarity or "world knowledge" on contemplated topics prior to test design (treating it therefore as an antecedent variable), or by attempting to provide some minimum topic information through the inclusion of pictures or graphic material. The facilitative role of visuals in language comprehension has been reviewed extensively by Levin and Lesgold (1978). Studies in writing performance have also attempted to control topic information by using pictures as writing stimuli (Pitts, 1978; Crowhurst, 1980). In addition, writing studies have found facilitative effects of pictures on writing coherence and support for lower achieving students (Baker & Quellmalz, 1980).

3) Task type/discourse mode. Task type appears to be a third task feature ignored in the design of test tasks. Cognitively oriented reading and writing studies are documenting that the recurring structural features of the separate modes of discourse, e.g., exposition, narration, argumenta-

tion, long described by rhetorical theorists, e.g., Kinneavy (1971), are extracted by readers and stored as schemata, or templates of regular discourse features. These schemata become "frames" for comprehending discourse. Research on learners' extractions of conventional discourse structures and applications of these schemata to new text has demonstrated that the schemata aids comprehension of narrative and expository discourse (Bransford & Johnson, 1972; Stein, 1978; Anderson et al., 1977; Meyer, 1975).

Writing research also provides a useful vehicle for understanding the importance of discourse mode in the general problem of test design. Studies have shown that writers employ different linguistic structures in different discourse modes and that writers are differentially skilled in producing essays in the modes (Quellmalz & Capell, 1979; Baker & Quellmalz, 1980; Praeter & Padia, 1980; Crowhurst, 1980). While researchers conjecture whether discourse conventions or schemata can or should be directly taught (Paris, Scardamaliz, & Bereiter, 1980), the cumulative evidence is that the differing structural features of discourse modes place quite different task demands on students. In discussing the syntactice shifts across discourse modes Cooper (1979) suggests that different modes will make a difference in product-oriented writing studies. He suggests, for example, that writing planning might change as much as 50 percent.

Also, in math learning, researchers have found similar performance sensitivity to task type and are exploring the unique and common operations within and between math task types. Davis and McKnight (1979) are studying the applications of frame or schema theory to mathematical learning. Brown is attempting to find common errors or "bugs" students make in relation to types of math tasks (Brown & Burton, 1978). In her discussion of generic, meta-cognitive skills, Hays-Roth (1980) describes one component which consists of a learner's repertoire of strategies and procedures for major problem types.

4) Structural complexity. A fourth task feature unexplicated in current domain specifications and highlighted in cognitive research is the structural complexity of passages. Work in discourse analysis (Kintsch, 1974; Meyer, 1975; Davis & Nold, 1978) indicates that semantic structure involving abstraction levels, amount of information, and number and type of relationships among pieces of information interacts with learners' comprehension, retention, and production of that information. Thus, designing reading math, or other subject-matter passages that fail to control for semantic structure could result in performance variability due to reader's comprehension difficulties with passage structure, rather than performance on the subject matter concepts and skills of interest. In addition, specifying rules for text structure would be more likely to guide production of somewhat homogeneous item pools.

5) Linguistic complexity. Cognitive learning research suggests a fifth content dimension, linguistic structure. The extensive literature documenting influence of linguistic complexity on comprehension supports the importance of linguistic control as an item difficulty or readability factor. (See, for example, Kintsch, 1974; Loban, 1976; Quillian, 1968.) Some reading test specifications attempt to control this through readability formulae (Fry, 1968; Klare, 1963); most test specifications, however, ignore the role of linguistic complexity in task design (see Polin & Baker, 1979).

Response Limits

Cognitive learning research has particular relevance for the specification of response limits in tasks. A marked failure of domain-referenced testing is inattention to the operations required for the student to react to the content. State-of-the-art test specifications tend to describe the response limits as constructed or selected, and specify the types of criteria or numbers of alternatives. It does not take a cognitive psychologist to realize that a greater range of different thought processing and problem solving goes into selecting multiple choice answers for some tasks compared to others.

1) Response mode. Learning research has long noted the performance distinction between selected and constructed response modes (Bourne, 1966; Skinner, 1954), but cognitive researchers are examining the operations required by the two response modes. In fact, much cognitive research elicits lengthy constructed responses from students during task processing, e.g., while solving an equation, while reading, or while planning, writing and revising, as well for the final response (the solution, the free recall, the essay). This pattern contrasts with the preference for multiple-choice options in actual tests.

Learning research in reading finds different information retained on cued (often multiple choice) vs. free recall tasks (Anderson et al., 1977). Writing process research is applying extensive protocol analyses to reports of writers' processes and to their productions elicited during planning, writing, and revising tasks, e.g., Hayes and Flower (1978, 1979), and Bereiter (1979). Writing assessment research in a domain-referenced framework has found very different descriptions of writing performance yielded by scores on multiple choice tests and writing samples (Spooner-Smith, 1978; Quellmalz & Capell, 1979; Spooner-Smith, Winters, Quellmalz & Baker, 1980). Math learning research finds that requiring production responses during problem solving and at the point of solution provides more diagnostic information about achievement (Davis, 1979; Brown & Vanlehn, 1980). In general, these researchers seem to value constructed responses in all subject areas as more reflective of the status of learner's skill development. This emphasis on constructed responding suggests that classroom level diagnostic tests, aimed at describing a student's competency status, should reduce dependence upon the conventional multiple choice mode. If tests can include constructed response tasks for "benchmark" stages of developing subject matter skill, instructors would have more powerful, sensitive diagnostic assessment devices.

2) Processing operations or routines. Cognitive learning research is beginning to test speculations about the nature and difficulty of operations involved in skilled performance. These studies on inferencing from text (e.g., Thorndyke, 1975; Spiro, 1975); routines involved in planning, writing, and revising (Shuy, 1977; Hayes & Flower, 1979); procedures for solving math problems (Brown & Burton, 1977; Davis & McKnight, 1979; Greeno, 1978); and operations in other subjects such as physics and chess (Simmon & Simon, 1978; Chi & Glaser, 1980; Larkin, 1979) aim to identify domain-relevant operations. Math research has identified routines for some skills (Davis & McKnight, 1979; Brown & Vanlehn, 1980; Resnick, 1980); reading and writing research are just beginning to identify and verify operations.

Error analyses of processing problems or "bugs" are being conducted to shed light on problem solving routines. Results from these analyses may not only inform test design about the routines required for different task types, but also about the types of distractors that could be generated for many multiple choice tasks. For example, Brown's research identifying common "bugs" or errors would suggest types of distractors to include on math multiple choice tests (Brown & Vanlehn, 1980). Shaughnessy, description of common writing errors might suggest classes of distractors for mechanics-oriented writing multiple choice tests (Shaughnessy, 1978).

3) Meta-cognitive strategies. A third dimension on which cognitive research suggests the response limits of test tasks can vary is on meta-cognitive, general reasoning skills. Cognitive researchers differ in their views on the place of these skills in subject matter test tasks. Royer and Cunningham (1978) propose that general reasoning skills are sepearate from, and should be removed from tests of reading comprehension. Brown, Campione, and Day (1980), however, define meta-cognitive information as that knowledge learners have about the state of their own knowledge base and strategies available to face task demands. They describe meta-cognitive strategies for approaching text comprehension which seem more specific to reading domains than to general strategies.

Hays-Roth (1980) defines meta-cognitive skills as strategies for knowledge acquisition, problem solving, and reasoning that are domain independent. These involve the learner's assessment of a repertoire of strategies suitable for a problem, selecting, scheduling, executing, and evaluating the efficacy of those strategies. Cognitive learning research with

with adults has correlated differences in meta-cognitive skills with differences in quality of learning and problem solving performance (Chi & Glaser, 1980; Hays-Roth, 1980). Frase (1980) also describes differences in test performance attributable to learners' inappropriate test taking strategies based on optional or partial representation of test tasks.

The research on meta-cognitive or executive strategies is clearly still exploratory. Its potential for test design may be the identification of learning and task engagement strategies important for students to activate across many subject domain tasks. Identification of generalizable strategies would permit either their assessment as antecedent variables or provision in test task design for partialing out or controlling for their relationsip to domain-specific operations. A line of current research in writing is concentrating upon identifying the set of meta-cognitive strategies required for competent writing (Flower, n.d.; Cooper & Matsuhashi, 1978; Rose, n.d.).

Summary of Implications of Behavioral and Cognitive Learning Research for Domain-Referenced Testing

This review of the implications of learning research for test design yields three major recommendations for design methodology. First, we have asserted that much state-of-the-art domain-referenced test development methodology could be vastly improved by simply improving the descriptive rigor of the content and response limits in domain-referenced test specifications. The content and response dimensions of test tasks should be described so clearly that they provide rules for replicably generating items with homogeneous content and response components. The content and response dimensions so explicated should be those that research suggests make a difference in student performance (and, therefore, items' conceptual and performance homogeneity).

Second, domain specifications should include those additional task features that research demonstrates influence learning and performance. These dimensions include content limits that affect examinees' understanding of the task demands: (1) context (e.g., purpose, audience, relevance, and time); (2) topic range (e.g., baseball games) and presentation mode (e.g., pictures); (3) discourse mode or task types (e.g., narration, exposition, three variable algebraic equations); (4) structure (e.g., inductive, deductive); and (5) linguistic complexity. Additional factors affecting examinees' responses should be included in response limits which should specify (1) response mode (e.g., direct, indirect) including scoring criteria; (2) required operations and distractor or discrimination parameters. Whether these specified task conditions do indeed differentially influence performance on the particular test can be empirically verified for a particular test.

Third, test design methodology should attempt to control for or tailor tasks to the antecedent conditions emerging as important learning and performance variables. Methods for dealing with antecedent factors include pretesting or tailoring tasks to individual's word knowledge, schemata or cultural, cognitive predispositions to engage test tasks in identifiable ways. It may be that constructed response tasks will permit identification of the interaction of student's entering repertoire with the task at hand.

Generally, as cognitive learning research documents the impact of antecedent or task conditions on student performance, test design methodology should more promptly and judiciously attempt to account or control for these conditions in assessment instruments. By attending to learning research, designers of tests (and instruction) can construct more valid, defensible, fair, and useful methodologies and measures.

References

Anderson, R. C. How to construct achievement tests to assess comprehension. Review of Educational Research, 1972, 42, 140-170.

Anderson, R. C. The notion of schemata and the educational enterprise. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), Schooling and the acquisition of knowledge. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977.

Anderson, R. C., Reynolds, R. E., Schallert, D. L., & Goetz, E. T. Frameworks for comprehending discourse. American Educational Research Journal, 1977, 14, 367-382.

Anderson, T. H., Wardrop, J. L., Hively, W., Muller, K. E., Anderson, R. I., Hastings, C. N., & Frederidsen, J. Development and trial of of a model for developing domain-referenced tests of reading comprehension. Urbana-Champaign, Ill.: Center for the Study of Reading, May, 1978.

Baker, E. L. Beyond objectives: Domain-referenced tests for evaluation and instructional improvement. Educational Technology, 1974, 14, 10-21.

Baker, E. L., & Quellmalz, E. S. Issues in eliciting writing performance: Problems in alternative prompting strategies. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, 1980.

Baker, E. L., & Spooner-Smith, L. Evaluation Response: Making Judgments. Los Angeles, CA: Center for the Study of Evaluation, Spring 1977.

Bartlett, F. C. Remembering. Cambridge, England: Cambridge University Press, 1932.

Bereiter, C. Development in writing. In Testing, teaching, learning. Report of a conference on research on testing. Washington, D. C.: National Institute of Education, 1979.

Bourne, L. E. J. Human conceptual behavior. Boston, MA: Allyn and Bacon, 1966.

Bransford, J. D., & Johnson, M. K. Contextual prerequisites for understanding. Some investigation of comprehension and recall. Journal of Verbal Learning and Verbal Behavior, 1972, 11, 717-726.

Britton, J., Burgess, T., Martin, N., McLeod, A., & Rosen, H. The development of writing abilities. New York: MacMillan Education, Ltd., 1975.

Brown, A. L., Campione, J. C., & Day, J. Learning to learn: On training students to learn from texts. Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

Brown, J. S., & Burton, R. R. Semantic grammar: A technique for constructing natural language interfaces to instructional systems. Cambridge, MA: Bolt, Beranek and Newman, Inc., 1977. (ERIC Document Reproduction Service No. ED 142 240)

Brown, J. S., & Burton, R. R. Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science, 1978, 2, 155-192.

Brown, J. S., & Vanlehn, K. Towards a generative theory of "bugs." Palo Alto, CA: Xerox Palo Alto Research Center, 1980.

Capell, F. Test design project: Studies in test bias, progress report. Center for the Study of Evaluation, University of California, Los Angeles, 1980.

Capell, F., & Quellmalz, E. Test design project: Studies in test bias. An examination of curricular relevance and curricular sensitivity of achievement tests in two languages. Los Angeles, CA: Center for the Study of Evaluation, 1980. (Grant No. NIE-G-80-1112)

Cazden, C. B. Two paradoxes in the acquisition of language structure and functions. In K. Connoly, & J. S. Bruner (Eds.), The growth of competence. New York: Academic Press, 1974.

Chi, M. T. H., & Glaser, R. The measurement of expertise: Analysis of the development of knowledge and skill as a basis for assessing achievement. In E. L. Baker, & E. S. Quellmalz (Eds.), Educational Testing and Evaluation - Design, Analysis, and Policy. Beverly Hills, CA: Sage Publications, 1980.

Chomsky, N. Syntactic structures. The Hague: Morton, 1957.

Cooper, C. R. Current studies of writing achievement and writing competence. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Cooper, C. R., & Matsuhashi, A. A video time-monitored observational study: The transcribing behavior and composing processes of a competent high school writer. Buffalo, NY: State University of New York, 1978.

Crowhurst, M. Syntactic complexity in narration and argument at three grade levels. Canadian Journal of Education, 1980.

Davis, B., & Nold, E. The discourse matrix. Palo Alto, CA: Stanford University, 1978. (ERIC Document Reproduction Service No. ED 168 022)

Davis R. B., & McKnight, C. C. Towards eliminating "black boxes": A new look at good vs. poor mathematics students. Urbana-Champaign, IL: University of Illinois, 1979.

Flavell, J. H. Cognitive development. Englewood Cliffs, NJ: Prentice-Hall, 1977.

Florio, S. Learning to write in the classroom community: A case study. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Flower, L. S. Good writing: Evaluating the writer's process. Pittsburgh, PA: Carnegie Mellon University, n.d. (mimeo).

Frase, L. J. The demise of generality in measurement and research methodology. In E. L. Baker, & E. S. Quellmalz (Eds.), Educational Testing and Evaluation: Design, Analysis, and Policy. Beverly Hills, CA: Sage Publications, 1980.

Fry, E. A readability formula that saves time. Journal of Reading, 1968, 7, pp. 513-516; 575-578.

Gagné, R. M. The conditions of learning (1st & 3rd ed.). New York: Holt, Reinhart and Winston, 1967, 1977.

Greeno, J. G. Cognitive objective in instruction: Theory of knowledge for solving problems and answering questions. In D. Klahr (Ed.), Cognition and Instruction. Hillsdale, NJ: Lawrence Erlbaum Associates, 1976.

Greeno, J. G. Nature of problem solving abilities. In W. K. Estes (Ed.), Handbook of Learning and Cognitive Processes (Vol. 5). Hillsdale, NJ: Lawrence Erlbaum Associates, 1978.

Hayes, J. R., & Flower, L. Protocol analysis of the writing process. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Hayes, J. R., & Flower, L. Writing as problem solving. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Hays-Roth, B. Cognitive skills for learning and thinking. Proposal submitted to the National Institute of Education, June 1980.

Herman, J., & Yeh, J. Test use: A review of the issues. In E. L. Baker & E. S. Quellmalz (Eds.), Educational Testing and Evaluation, Beverly Hills, CA: Sage Publications, 1980.

Hively, W. Introduction to domain-referenced testing. Educational Technology, 1974, 14, 5-10.

Kinneavy, J. R. A theory of discourse. Englewood Cliffs, NJ: Prentice-Hall, 1971.

Kintsch, W. The representation of meaning in memory. Hillsdale, NJ: Lawrence Erlbaum Associates, 1974.

Klare, G. R. The measurement of readability. Ames, Iowa: Iowa State University Press, 1963.

Larkin, J. H. Skilled problem solving in physics: A hierarchical planning approach. Journal of Structural Learning, 1979.

Levin, J. R., & Lesgold, S. N. On pictures in prose. Educational Communication and Technology, 1978, 26, 233-243.

Loban, W. Language development: Kindergarten through grade twelve. Urbana, IL: National Council of Teachers of English, 1976.

Meyer, B. F. The organization of prose and its effects on memory. North Holland studies in theoretical poetics (Vol. 1). Amsterdam: North Holland Publishing Company, 1975.

Minsky, M. A framework for representing knowledge. In P. H. Winston (Ed.), The psychology of computer vision. New York: McGraw-Hill, 1975.

Montague, W. E. A common flaw in research design: Inconsistency between learning and testing requirements. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

Nessier, U. Cognitive psychology. New York: Appleton-Century-Crofts, 1967.

Nold, E. W. The writing process. Unpublished manuscript. Palo Alto: Stanford University, 1979.

Norman, D. A., & Bobrow, D. G. On data-limited and resource-limited processes. Cognitive Psychology, 1975, 7, 44-64.

Norman, D. A., & Rumelhart, D. E., & the LRN Research Group. Explorations in cognition. San Francisco, CA: Freeman, 1975.

Olson, D. From utterance to text: The bias of language in speech and writing. In H. Fisher, & R. Diez-Guerro (Eds.), Language and Logic in Personality and Society. New York: Academic Press, 1977.

Paivio, S. Imagery and verbal processes. New York: Holt, Rinehart & Winston, 1971.

Paris, P., Scardamalia, M., & Bereiter, C. Discourse schemata as knowledge and as regulators of text production. Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

Pitts, M. The relationship of classroom instructional characteristics and writing in the descriptive/narrative mode. Report to the National Institute of Education. Los Angeles, CA: Center for the Study of Evaluation, November 1978 (Grant No. OB-NIE-G-78-0213).

Polin, L. G., & Baker, E. L. Qualitative analysis of test item attributes for domain-referenced content validity judgments. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Popham, W. J. Domain-referenced strategies. In R. A. Berk (Ed.), Criterion-referenced measurement. Johns Hopkins University Press, 1980.

Praeter, D., & Padia, W. Effects of modes of discourse in writing performance in grades four and six. Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

Quellmalz, E. S. Test design: Aligning specifications for assessment and instruction. Paper presented at the conference Evaluation in the 80's: Perspectives for the National Research Agenda. Los Angeles, CA: Center for the Study of Evaluation, June 1980.

Quellmalz, E. S., & Capell, F. Defining writing domains: Effects of discourse and response mode. Report to the National Institute of Education. Los Angeles, CA: Center for the Study of Evaluation, November 1979. (Grant No. OB-NIE-G-78-0213)

Quellmalz, E. S., & Snidman, N., & Herman, J. Toward competency-based reading systems. Paper presented at the annual meeting of the American Educational Research Association, New York, 1977.

Quillian, M. R. Semantic memory. In M. Minsky (Ed.), Semantic information processing. Cambridge, MA: MIT Press, 1968.

Resnick, L. What do we mean by meaningful learning? Invited address at the annual meeting of the American Educational Research Association, Boston 1980.

Royer, J. M., & Cunningham, D. J. On the theory and measurement of reading comprehension (Tech. Rep. No. 91). Urbana-Champaign, IL: Center for the Study of Reading, 1978.

Rose, M. Strategies, audience, exposition and freshman's process - A cognitive/contextual theory of instruction for college composition. Los Angeles: University of California, Writing Research Project, n.d.

Rudner, L. M. A short and simple introduction to tailored testing. Paper presented at the Eastern Educational Research Association annual meeting, Williamsburg, VA, March, 1978.

Rumelhart, D. E., & Ortony, A. The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), Schooling and the Acquisition of Knowledge. New York: Lawrence Erlbaum Associates, 1977.

Scribner, S., & Cole, M. Unpackaging literacy. Social Science Information, 1978, 17, 19-49.

Shaughnessy, M. P. Errors and Expectations. New York: Oxford University Press, 1977.

Shuy, R. W. Toward a developmental theory of writing: Tapping and knowing. Paper presented at the National Institute of Education Conference on Writing, Los Alamitos, CA, 1977.

Simon, D. P., & Simon, H. A. Individual differences in solving physics problems. In R. Siegler (Ed.), Children's thinking: What develops? Hillsdale, NJ: Lawrence Erlbaum Associates, 1978.

Skinner, B. F. The science of learning and the art of teaching. Harvard Educational Review, 1954, 24, 86-97.

Skinner, B. F. Verbal behavior. New York: Appleton-Century-Crofts, 1957.

Solomon, D. Intergroup relations in the elementary school: The effects of classroom environments. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

Spiro, R. J. Inferential reconstruction in memory for connected discourse. (Tech. Rep. No. 2) Urbana-Champaign, IL: Cognitive Studies in Education, October, 1975.

Spooner-Smith, L. Investigation of writing assessment strategies. Report to the National Institute of Education. Los Angeles, CA: Center for the Study of Evaluation, November 1978. (Grant No. OB-NIE-G-78-0213)

Spooner-Smith, L., Winters, L., Quellmalz, E., & Baker, E. L. Characterizations of student writing competence: An investigation of alternative scoring systems. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

Stallard, C. An analysis of the writing behavior of good student writers. Research in the Teaching of English, 1978, 12, 206-218.

Stein, N. L. How children understand stories: A developmental analysis (Tech. Rep. No. 69). Urbana-Champaign, IL: Center for the Study of Reading, 1978.

Thorndyke, P. W. Cognitive structures in human story comprehension and memory (Tech. Rep. No. P5513). Santa Monica, CA: Rand Corp., 1975. (ERIC Document Reproduction Service No. ED 123 587)

Wardrop, J. L., Anderson, T. H., Hively, W., Anderson, R. I., Hastings, C.N., & Muller, K. E. A framework for analyzing reading test characteristics. Urbana-Champaign, IL: University of Illinois, 1978.

Wittrock, M. C. Learning as a generative process. Educational Psychology, 1974, 11, 87-95.

EFFECTS OF DISCOURSE AND RESPONSE MODE ON THE MEASUREMENT OF WRITING COMPETENCE

Edys Quellmalz, Frank J. Capell,
and Chih-Ping Chou

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, CA 90024

Grant No. OB-NIE-G-80-0112

The research reported herein was supported in whole or in part by a grant to the Center for the Study of Evaluation from the National Institute of Education, U.S. Department of Education. However, the opinions and findings expressed here do not necessarily reflect the position or policy of NIE and no official NIE endorsement should be inferred.

EFFECTS OF DISCOURSE AND RESPONSE MODE ON THE MEASUREMENT OF WRITING COMPETENCE

As school district and state assessment programs attempt to test student basic skills achievement, attention to the methodological problems inherent in measuring writing competence increases. The complexity of writing as a skill domain and the lack of consensus about its components have engendered much controversy about the important characteristics of writing tasks. There is little agreement about the type, length or number of tasks that should be administered for a given test form and even about whether some aspects of composition require "direct" assessment through the elicitation of writing samples.

Two salient measurement issues involved in specifying writing task types are the response mode (selected vs. constructed) and the discourse mode required by the tasks. Conventionally, large scale assessments have dealt with the response mode issue by measuring writing skills indirectly with multiple choice tests. Support for such indirect measurement derived from reported high correlations between objective tests and direct measures of written production, from the erratic reliabilities accompanying impressionistically scored essays and from the economic and logistical demands of collecting and scoring writing samples (Braddock, Lloyd-Jones, & Schoer, 1963; Godshalk, Swineford & Coffman, 1966; Breland, 1977). More recently, however, demands for writing tests with content, construct and ecological validity have prodded reinstatement of direct, written production tasks.

Yet even when assessments collect writing samples, students usually produce only one composition, despite the well documented fluctuation of writing performance from one sample to the next (French, 1962; Braddock et al., 1963; Diederich, 1974). By necessity, a single sample taps student performance on only one type of discourse and on one topic. This limitation presents a measurement problem since the generic methods of development in particular forms of discourse differ substantially from one another. Salient structural features of argument, for instance, are issues, reasons, and conclusions, while stories include plot, character, setting, and theme. Exposition involves main idea, supporting detail and logical development; narrative elaborates events in chronological order, and description portrays concrete details in spatial order. Since different purposes set for writing tend to elicit the generic structural elements of the modes of discourse (Kinneavy, 1971), it seems likely that the schema or frames activated in the writer by writing tasks varying in purpose should differ (Anderson, 1977; Minsky, 1975).

Research on Discourse Mode Effects

Evidence from reading and writing research support the distinctiveness of processes required by varying discourse modes. Reading research suggests that different schema are used as students attempt to comprehend narrative and expository text (Meyer, 1975; Graesser, Hauft-Smith, Cohen & Pyles, 1979). As in reading, writers also employ various skills and personal resources to meet the demands of a kind of writing or mode of discourse. For students learning to write, these different discourse modes represent very dissimilar challenges and, furthermore, attempts to compare student writing skills across modes of discourse constitute a real assessment issue. In one of the most frequently cited writing assessment studies, the performance variability in "topic" discussed by Godshalk, et al. was, probably, equally or more attributable to the five different discourse modes stimulated by the assignments than to the subject matters addressed. Veal and Tillman (1971) reported variability in elementary students' performance on tasks specifying different discourse aims as did Praeter and Padia (1980).

Moreover, other writing research demonstrates that different writing purposes stimulate writing that varies in structural complexity (Crowhurst & Piche, 1979; Crowhurst, 1980; Perron, 1977) and that represents writing topics quite differently (San Jose, 1973; Perron, 1977). Most importantly for instruction and evaluation, this accumulating body of writing research suggests that different writing purposes require dissimilar writing strategies of varying difficulty for individual students. Cooper (1979) has cited research indicating that sentence structures shift when discourse mode changes and speculated that a student's planning demands for an essay might change as much as 50%.

The implications of reading and writing studies on discourse mode effects for writing assessment are that the mode of discourse of the writing purpose will make a difference in writing performance. For examples, students might be more skilled at narrative writing tasks requiring chronological development than at expository tasks requiring logical development. Thus the profile of writing competence for a student derived from a writing test calling for exposition may differ from the profile of writing competence for that same student derived from a narrative or persuasive task.

Research on Response Mode Effects

In addition to the question of skill commonality across discourse modes or genres, the question of the response mode or measurement form in which the writing skill should be assessed continues as a hotly debated topic. While many claims are made for the predictive or concurrent validity of indirect, objective writing measures (Coffman, 1971; Breland, 1977), indirect measures simply are not considered by writing researchers to meet the more crucial standards of content or construct validity (Braddock, et al., 1963; Cooper & Odell, 1977). Learning theory and research suggests that selected responses elicited by multiple choice tests provide some valuable information, but are, after all, measures of processes required in reading comprehension, not measures of actual production ability. As such, recognition tasks are often considered, at best, behaviors enroute to constructed responses (Bourne, 1966, Skinner, 1963).

Comparisons of direct and indirect writing measures have yielded moderate correlations between the scores resulting from the two response modes. In one of the seminal writing assessment studies, Godshalk, et al. (1966) reported correlations from .46 to .75 between the sum of 5 essay scores from high school students and their College Board English Comprehension Test. In an attempt to validate ETS's Test of Standard Written English (TSWE) Breland, Conlan and Ragosa (1976) found correlations of only .42 between that mechanics-oriented test and a 20 minute essay scored on a 4-point scale. In a subsequent study, Breland and Gaynor (1979) reported correlations ranging from .58 to .63 between students' three separate essay scores awarded on a 6-point scale and the TSWE, while the correlation between the sum of the three essays and the TSWE was .76. Similar low to moderate correlations of .43 to .67 were found in a comparison of the American College Testing Program's English Usage Test (also emphasizing sentence-level skills) and students' scores on three essays. Hogan and Mishler's (1980) study of the relationship between third and eighth grade students' Metropolitan Achievement Test scores and one essay yielded correlations of .68 and .65, while the correlations increased to .75 and .81 when a second essay score entered into the calculations.

In general, the methodology of these studies has related essay scores derived from norm-referenced holistic ratings to a total multiple choice test score. The apparent assumption was that both sets of measures tapped the same set of writing skills. However, content analyses of the essay rating criteria reveals that they were often vaguely worded, but did reference whole-text features such as thesis, coherence, support and style, as well as sentence-level mechanical conventions. Items on the multiple choice tests, on the other hand, often emphasized sentence-level mechanics and required few if any text-level discriminations and, obviously, no production responses. At issue then is not simply whether measures correlate statistically, but whether different measures focus on the same text features of written productions and whether they reflect the same underlying skill constructs.

Design Requirements for Comparing Alternative Measures of the Same Construct

To compare the information yield and psychometric quality of writing measures involving different discourse and response modes, data are needed that contrast the performance of a group of examinees across equivalently specified skill domains varying only by the modes of measurement. The specific test objectives (skill domains), stimulus dimensions, instructions to examinees, and response criteria/characteristics need to be matched as closely as possible across discourse and response modes; i.e., each of the measures should present parallel content-valid procedures for assessing the same skill or skills. Data from measures designed to be psychologically parallel can then be examined in terms of the comparative construct validity and reliability of the discourse and response mode variables. A test of writing researchers' contentions that text-level writing skills such as thesis statement, organizationation and support are best measured by written production tasks would involve comparing multiple choice test subscale scores measuring comprehension of passages for such subskills as organization, support, and mechanics with ratings of these features in text the student produces. This paper reports such comparisons for score profiles obtained from analytically scored direct assessments of student writing (essay and paragraph length writing samples) and an indirect assessment (multiple choice questions concerning prose passages). Measures were designed to be conceptually parallel by using the domain specifications that guided production of directions/prompts for the writing tasks to construct the passages employed in the multiple choice task. Similarly, the dimensions of writing quality making up the analytical scoring rubric applied to the writing samples determined the specific aspects of the prose passages the multiple choice measure questioned.

The measurement issues addressed in the study concerned the comparability of writing scores obtained from tasks varying in discourse and response mode. The study departed from more conventional methodology in two respects. First, the measures of writing skills in the different response modes were specifically designed to present tasks parallel on all dimensions but the discourse and response mode variables. Second, the study augmented the standard correlational comparisons of the measures with a multitrait multimethod (MTMM) analytical technique designed to identify the factor structure underlying student's writing score profiles derived from the alternative measures. These analyses examined the variance of the writing scores derived from different discourse and response modes and the comparative discriminant validity or distinctiveness of the sets of scores across mode variations, treating the scale scores comprising the writing profiles as "traits," and the discourse or response modes as "methods" (Campbell & Fiske, 1957). Correlations among different operationalizations of the same variable should arise from the influence of a single common factor or trait (e.g., organization). Also the method of measurement should exert an influence on each variable, so that variables measured by a common method will covary to a greater degree than those measured by different methods; this covariation can be thought of as reflecting the operation of a common method factor. This formulation of the MTMM approach has been implemented in other empirical studies through confirmatory factor analysis (Jöreskog, 1974; Traub & Fisher, 1977; Werts, Jöreskog & Linn, 1972). Traub and Fisher (1977) for example, compared verbal and quantitative scores derived from fill-in, right/wrong multiple choice and partial knowledge multiple choice response formats in exactly this fashion.

Two studies attempted to compare direct and indirect measures of reasonably parallel text features. Spooner-Smith used domain-referenced skill specifications to design multiple choice items analogous to essay rating criteria. She found correlations of the multiple choice test total score with a General Impression score of .65 and with the total of analytic ratings of .61. Relationships between analogous features such as Organization and Support, however, were much lower, ranging from .23 to .55. Her findings suggested that when multiple choice scores and essay scores derived from precisely matched definitions of text features, the comparability of scores on these component writing skills might be even lower than those previously reported.

In this study we asked 1) whether student writing performance profiles are comparable for tasks differing in discourse mode (writing purpose), and 2) whether tasks requiring different response modes (paragraphs, essays, and multiple choice items) provide the same type and quality of information about student writing competence. In the MTMM framework, we examined whether distinctive common factors underlay the corresponding variables from the writing profiles derived from the discourse and response modes variations.

Method

To examine the relationship of writing scores yielded by tasks differing on the two variables, discourse and response mode requirements, high school students received writing tests on three separate occasions. Each student received a multiple choice test and a paragraph writing task as well as two full length essay assignments. Ratings of the essays and paragraph on an analytic scale and scores on the objective test provided the bases for the comparisons.

Sample

Approximately two hundred eleventh and twelfth grade students from three high schools in a small school district in the Los Angeles area participated in the study. Students were selected who were attending English or composition classes that were judged by teachers to contain average or above average pupils. Scores from the verbal portion of the Differential Aptitude Test were available for 92 students in the sample; the mean percentile score for this subsample was 63.9 (s.d. = 28.6).

Design

Students within each class were randomly assigned to one of four testing conditions defined by the relationship of the discourse mode(s) of the essay tasks. In Conditions 1 and 2 (Same Genre), the three constructed response writing tasks (two essays and one paragraph) were in the same discourse mode. Condition 1 students wrote two expository essays and an expository paragraph; Condition 2 students wrote two narrative essays and a narrative paragraph.

In Conditions 3 and 4, (Different Genre) students wrote one narrative and one expository essay. Condition 3 students wrote an expository essay on Topic A and a narrative essay on Topic B, while Condition 4 students wrote an expository essay on Topic B and a narrative essay on Topic A. Half of the subjects in Conditions 3 and 4 wrote an expository paragraph, while half wrote a narrative paragraph.

Response mode, the second factor, was a within-subject factor and consisted of the multiple choice test (selected response), the paragraph (short constructed response) and the essay (long constructed response). During the three testing occasions subjects received the multiple choice test and paragraph on one occasion, and an essay on each of the other two occasions. The design counterbalanced the order in which students received the tasks.

Measures

The essay and paragraph tasks were constructed in accordance with a set of domain specifications for expository and narrative writing. These specifications included the purpose of the writing assignment, guidelines for appropriate topics, the response criteria by which written products were to be judged, and guidelines for the content and format of the directions for the tasks. The response criteria were chosen to reflect the discourse features of an analytic scoring system developed at UCLA (Pitts, 1978; Spooner-Smith, 1978; Winters, 1978; Quellmalz, 1979). The version of the scoring system used in this study generated five ratings for each written product:

(1) General Impression -- A global judgment of writing quality assigned by raters after a quick initial reading of the writing sample.

(2) Focus -- The extent to which the subject and main idea of the writing sample were clearly stated or implied.

(3) Organization -- The extent to which the main idea was developed according to a discernible method of organization (e.g., clear chronological or logical development).

(4) Support -- The extent to which generalizations and assertions were supported by specific, relevant, subordinant statements.

(5) Mechanics -- The extent to which the writing sample was free from intrusive sentence-level mechanical errors (i.e., usage, sentence construction, spelling, capitalization and punctuation).

Each essay and paragraph was assigned ratings on these five subscales by one of two pairs of trained raters, the median generalizability coefficients for the two rater pairs were .61 and .83 across topics/occasions and subscales. The three writing samples representing direct measurement (two essays and one paragraph) generated 15 subscale scores, each on a one (low) to four (high) scale. The scores were calculated by averaging the scores assigned by both raters to each essay and paragraph for each subscale.

The stimulus attributes from the specifications for the writing tasks were used to develop the passages to be read in the multiple choice task. Ten passages were constructed, five expository and five narrative. For each passage, there were three questions, designed to be analogous to text features included in the rating scales -- main idea (focus), organization, and support. Main idea questions were referenced to a stated generalization near the beginning or end of the passage. Organization questions required the selection of a new sentence that would best fit at a point in the passage marked by an arrow. Support questions asked which new sentence would best support the main idea of the passage.

Results

Discourse Mode Effects

The first set of analyses compared students' scores according to the discourse mode of the task.

Table 1 presents means and standard deviations of essay ratings for each of the four test conditions.

- - - - - - - - - -

Insert Table 1 here

- - - - - - - - - -

On all five subscales and on total essay scores, narrative ratings were lower than expository ratings. This finding may be due to the differential curricular emphasis given to narrative and expository writing in the high schools, to subjects' lack of knowledge at a personal experience level with information required to deal with the narrative topics, or to raters' tendency to score narratives more stringently.

Table 2 displays the correlations between students' two essay scores on each of the analytic scale subscales. As expected, correlations between essay scores of students writing two essays in the same discourse mode (Same Genre) are higher than those of students in the Different Genre conditions.

- - - - - - - - - -

Insert Table 2 here

- - - - - - - - - -

An examination of individual subscales across conditions suggests that General Impression and Organization seem to differentiate most between the different discourse modes. This finding might also be expected,

Table 1

Means and Standard Deviations of Essay Scores

Test Condition		Same Genre				Different Genre			
		1		2		3		4	
Topic		A	B	A	B	A	B	A	B
General Impression	X̄	2.20	2.01	1.27	.88	1.43	1.89	.94	2.09
	sd	.60	.70	1.44	1.09	1.89	.66	.88	.54
Focus	X̄	2.45	2.30	2.21	2.21	2.28	2.19	2.26	2.33
	sd	.69	.64	.68	.64	.59	.69	.51	.53
Organization	X̄	2.16	1.98	1.88	1.62	1.88	1.95	1.52	2.08
	sd	.72	.70	.98	.71	.85	.70	.56	.56
Support	X̄	2.34	2.38	2.42	2.19	2.36	2.28	2.10	2.26
	sd	.64	.54	.81	.66	.70	.65	.54	.53
Mechanics	X̄	2.35	2.34	2.56	2.42	2.28	2.03	2.35	2.42
	sd	.64	.68	.69	.67	.80	.81	.56	.54
Total	X̄	11.50	10.90	10.35	9.32	10.23	10.31	9.18	11.19
	sd	2.78	2.67	3.84	2.89	3.36	3.02	2.02	2.05
n =		40	40	39	39	40	40	54	54

Table 2

Correlation Between Students' Two Essays

	Same Genre		Different Genre	
Condition	1	2	3	4
General Impression	.56	.39	.33	-.08
Focus	.43	.68	.41	.37
Organization	.42	.42	.20	-.10
Support	.27	.28	.07	.38
Mechanics	.58	.68	.63	.50
Total	.60	.55	.41	.22
n =	40	39	40	54

since General Impression requires a judgment about the global quality of the essay as an example of exposition or narration. Therefore the constellation of essay factors influencing this judgment should be the most comprehensive for each discourse mode and thus the most discriminating. Structurally, exposition and narration differ dramatically in their characteristic use of logical or temporal organizations, respectively. On the Mechanics subscale, correlations across conditions are most comparable, reinforcing the notion that the constellation of syntactic, punctuation, spelling and usage skills may not vary between modes of discourse so much as text-level skills do.

To test the statistical difference between students' scores received in the same and different discourse mode conditions, the correlations in Table 3 were transformed to standardized scores, average standardized scores were calculated for the same and for different genre conditions, and these average standardized scores for same and different conditions were contrasted. Table 3 presents the results of this analysis.

- - - - - - - - - -

Insert Table 3 here

- - - - - - - - - -

The comparison reveals that the realtionship between student's two essay scores on General Impression, Organization, and the Total is significantly stronger when students write in the same genre than when they write in different genre.

Comparisons between discourse modes on the paragraphs were conducted only at the group level, since an individual student wrote only one paragraph. Table 4 presents the results of these comparisons. The analytic

Table 3

Comparison Between Essay Scores
Same and Different Genre Testing Conditions

	Same Genre			Different Genre			
Condition	1	2		3	4		
	z	z	$z_{ave.12}$	z	z	$z_{ave.34}$	$Z_{observed}$
General Impression	.633	.412	.5225	.343	-.08	.1315	2.516*
Focus	.460	.829	.6445	.436	.388	.412	1.496
Organization	.448	.448	.448	.203	-.100	.0515	2.551*
Support	.277	.288	.2825	.070	.400	.235	.306
Mechanics	.662	.829	.7455	.741	.549	.645	.647
Total	.693	.618	.6555	.436	.224	.33	2.095*
n =	40	39	79	40	54	94	

$^*p<.05$

scales for narration and exposition were used for rating the paragraphs.

- - - - - - - - - - - - -

Insert Table 4 about here

- - - - - - - - - - - - -

Subscale scores ranged from 1-4, total scores from 1-20. Ratings of narrative paragraphs were generally lower than ratings of expository paragraphs. Ratings of expository paragraphs differed significantly from narrative paragraphs on the General Impression, Focus, and Organization subscales and on the Total scores, suggesting that as a population, these high school students wrote more skillfully in the expository mode. Consonant with essay data, Mechanics and Support were not as influenced by the different discourse tasks.

Multiple choice test comparisons of interest were the scores each individual received on the narrative and expository sections of the exam. Table 5 presents the means and standard deviations.

- - - - - - - - - - - - -

Insert Table 5 about here

- - - - - - - - - - - - -

On this reading comprehension test of items measuring recognition of writing-related skills, students were able to answer Focus/Main Idea and Support questions similarily well for both expository and narrative passages. On Organization questions however, students had more difficulty in general (73% overall average), particularly with narrative organization (66%).

Table 6 displays the correlations between individuals' expository and narrative scores on the three multiple choice subscales.

Table 4

Differences Between Scores on Expository and Narrative Paragraphs

		Expository Paragraph	Narrative Paragraph
General Impression	$\overline{X}$	1.93	1.05
	s.d.	.71	1.06
	t =	6.74***	
Focus	$\overline{X}$	2.34	2.09
	s.d.	.54	.63
	t =	2.93**	
Organization	$\overline{X}$	1.95	1.70
	s.d.	.62	.78
	t =	2.41**	
Support	$\overline{X}$	1.94	2.02
	s.d.	.67	.71
	t =	.87	
Mechanics	$\overline{X}$	2.35	2.29
	s.d.	.65	.74
	t =	.65	
Total Score	$\overline{X}$	10.50	9.15
	s.d.	2.66	2.93
	t =	3.37**	
	n =	111	n = 89

Table 5

Means and Standard Deviations of Multiple-Choice Test

		Expository	Narrative	Total
Focus	$\overline{X}$	4.61 (92%)	4.50 (90%)	9.13 (91%)
	s.d.	.77	.89	1.38
Organization	$\overline{X}$	4.05 (81%)	3.31 (66%)	7.39 (73%)
	s.d.	1.03	1.13	2.03
Support	$\overline{X}$	4.56 (91%)	4.41 (88%)	8.97 (90%)
	s.d.	.77	.98	1.51
Total	$\overline{X}$	13.23 (88%)	12.23 (82%)	25.33 (84%)
	s.d.	2.02	2.43	4.26

n = 241

- - - - - - - - - - -

Insert Table 6 here

- - - - - - - - - - -

While all correlations are statistically significant, the correlations between the narrative and expository subscales are substantially lower ($r=.47$ to .49) than the correlations between the total expository and narrative scores ($r = .65$).

In the aggregate, the preceding analyses constrasting students' performance on writing tasks differing in discourse mode suggest that: 1) students' writing skills vary in the different discourse modes, and 2) discourse mode score variability seems to be differentially distributed across writing subskills. These results occurred in separate analyses of students' essays, paragraphs and multi-choice scores.

To test the effect of discourse mode on subskills, the data were then subjected to multi-trait multi-method (MTMM) analyses using confirmatory factor analysis techniques (Jöreskog, 1974). Traits were defined as the writing subskills; methods were defined as exposition and narration. The analyses required within-subject measures of trait and discourse mode, therefore data from 94 subjects in test conditions 3 and 4 were used to examine the effects of discourse mode. In these conditions students wrote an expository and narrative eessay and answered multiple choice questions about expository and narrative passages. Since students wrote only one paragraph, paragraph scores were not included in the analyses.

To examine the factor structure of discourse modes in essay performance, the five analytic subscales of General Impression, Focus, Organization, Support and Mechanics formed the trait dimension. These five trait scores were composed by averaging scores over raters and standardizing across

Table 6

Correlations Between Expository and Narrative Multiple Choice Scores

	Expository Subscales				
Narrative Subscales	Focus	Organization	Support	Expository Total	Test Total
Focus	.47	.32	.41	.50	.64
Organization	.30	.48	.33	.49	.73
Support	.43	.35	.49	.53	.64
Narrative Total	.50	.52	.51	.65	.88
Test Total	.64	.73	.63	.86	

n = 241

topics. Thus ten scores were constructed, two each (expository and narrative) for each of the five subscales.

To examine the factor structure of discourse modes across response modes, a second set of analyses used just the three subcales (traits) common to the essay and multiple choice tests, Focus, Organization, and Support. For the multiple choice test, number correct scores were formed within discourse mode for each of the three multiple choice subscales. A trait score for each discourse mode was then formed by standardizing scores across the two response modes (essay and multiple choice). The standardization was aimed at removing interactions between response mode and genre. This second set of analyses employed six scores, two each (expository and narrative) for Focus, Organization and Support. Table 7 describes the variables and their abbreviations.

- - - - - - - - - - - - - -

Insert Table 7 about here

- - - - - - - - - - - - - -

All analyses were based on correlation matrices computed for the variables in Table 7. Maximum likelihood estimates of the parameters of the MTMM confirmatory factor analysis models were obtained from the LISREL computer program for the analysis of covariance structures (Jöreskog, 1973, 1977; Jöreskog & Sorbom, 1978). The LISREL program allows the analyst to treat model parameters (e.g., factor loadings or factor intercorrelations) in one of three ways: (a) as free parameters to be estimated by the program; (b) as fixed parameters specified in advance to equal some fixed number (usually zero); or (c) as constrained parameters to be estimated by the program subject to the constraint that they equal other estimated

TABLE 7

Description of Variables in LISREL Analyses of Discourse Mode Effects

Analyses For Essay Data

Trait Variables	Method Variables Expository	Narrative
General Impression	GIE	GIN
Focus	FI	FN
Organization	OE	ON
Support	SE	SN
Mechanics	ME	MN

Analyses Pooling Essay and Multiple Choice Data

Trait Variables	Method Variables Expository	Narrative
Focus	FE	FN
Organization	OE	ON
Support	SE	SN

parameters. In addition, the program computes standard errors for all free and constrained parameters, as well as an overall chi square test of the model's fit to the data. All model equations (in LISREL notation) are of the form:

$$\underset{\sim}{Y} = \underset{\sim}{\Lambda}_Y \underset{\sim}{\zeta} + \underset{\sim}{\varepsilon} \qquad (1)$$

$$\underset{\sim}{\Sigma}_Y = \underset{\sim}{\Lambda}_Y \underset{\sim}{\Psi} \underset{\sim}{\Lambda}'_Y + \underset{\sim}{\theta}_\varepsilon \qquad (2)$$

Each observed score in $\underset{\sim}{Y}$ depends on the latent variables $\underset{\sim}{\zeta}$ and $\underset{\sim}{\varepsilon}$, common factors and measurement errors, respectively. Equation (2) shows the hypothesized structure underlying the covariance matrix of the Y's; it consists of a matrix of factor loadings $\underset{\sim}{\Lambda}_Y$ (hereafter denoted Lambda), the covariance matrix of the $\underset{\sim}{\zeta}$'s, $\underset{\sim}{\psi}$ (hereafter Psi), and a (usually) diagonal matrix of error variances. Interest in the analyses to follow focuses on the contents of Lambda and Psi.

The first set of analyses examined the influence of discourse mode on writing trait scores derived from essays. The correlation matrix of the transformed scores appears in Table 8. Figure 1 displays a path diagram for a model specifying the five trait subscale factors and the two discourse modes/method/factors.

- - - - - - - - - - -

Insert Table 8 here

- - - - - - - - - - -

- - - - - - - - - - - -

Insert Figure 1 here

- - - - - - - - - - - -

Each test score was assumed to be affected by a discourse mode, either narrative or expository, and a writing subskill. Since writing skill might be affected by other factors like I.Q. or reading ability, the five traits

TABLE 8

Correlation Matrix of Scores Containing Discourse Mode and Subscale Trait Effects

	GIN	GIE	FN	FE	ON	OE	SN	SE	MN	ME
GIN	1.000									
GIE	.380	1.000								
FN	.608	.308	1.000							
FE	.306	.523	.327	1.000						
ON	.672	.174	.690	.330	1.000					
OE	.087	.502	.108	.456	.007	1.000				
SN	.609	.301	.474	.060	.600	.116	1.000			
SE	.230	.571	.066	.461	-.016	.478	.087	1.000		
MN	.411	.482	.351	.222	.286	.125	.322	.297	1.000	
ME	.198	.642	.083	.386	.015	.354	.109	.410	.545	1.000

N = narrative

E = expository

GI = General Impression

F = Focus

O = Organization

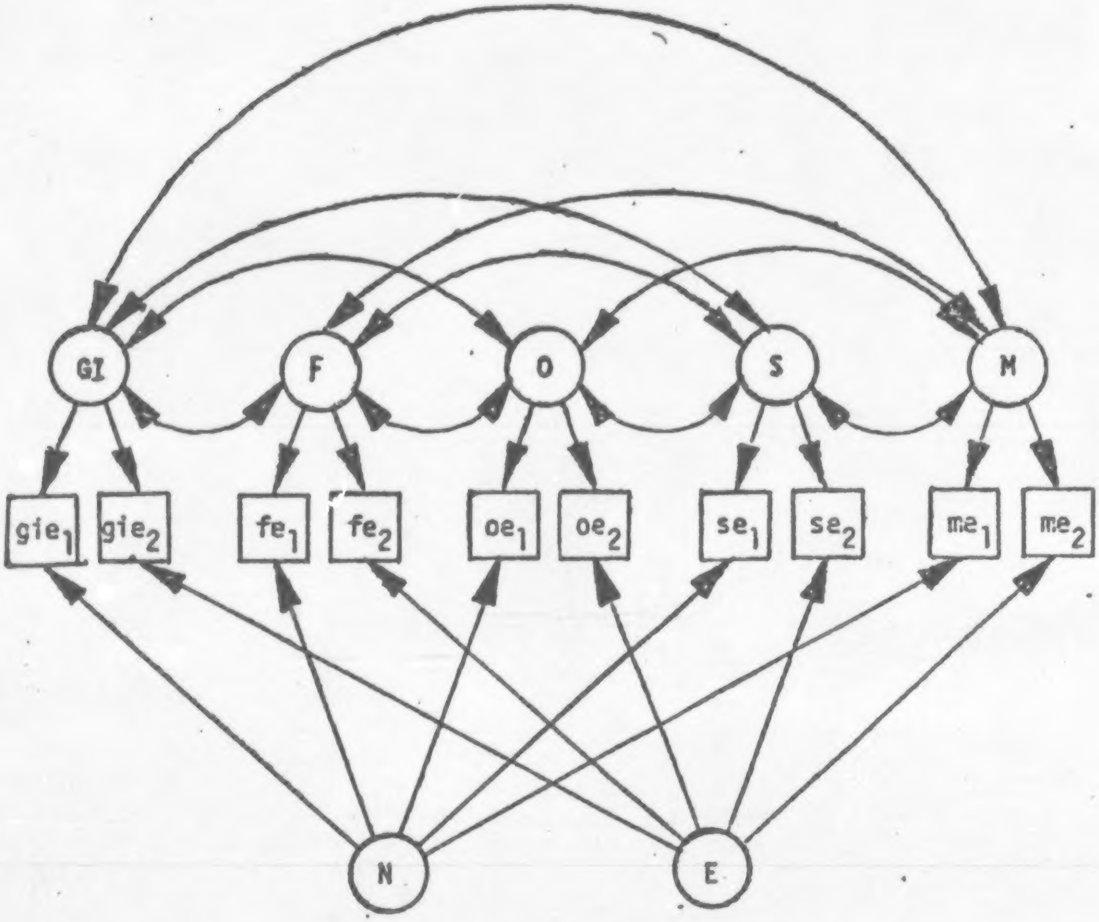
S = Support

M = Mechanics

FIGURE I

Discourse Mode

Path Diagram for Model I


GI
F
O
S
M
gie1
gie2
fe1
fe2
oe1
oe2
se1
se2
me1
me2
N
E

or writing skills were assumed to be distinctive, but interrelated, while the two discourse modes were assumed to be independent. With the effect of each trait constrained to be the same on the two discourse mode test scores, the estimates of the parameters and their standard errors (in parentheses) are summarized in Table 9.

- - - - - - - - - -

Insert Table 9 here

- - - - - - - - - -

The chi-square value of 19.94 yields a non-significant probability of .46. This result indicates that the observed test scores fit this model adequately from a global point of view.

Examination of all the LISREL estimates suggests that a better model might be formed to fit the data. Loadings of test scores on their corresponding traits range from .194 for the Organization subscale to .726 for the Mechanics subscale. The test of significance for the loading of each observed score on trait shows that traits have a significant effect on their corresponding test scores with the exception of the trait of Organization. The correlation among all traits, or psi-coefficients, indicates that three coefficients are abnormally high.[1] They are the correlations between General Impression and Organization, between General Impression and Support, and between Focus and Organization, as indicated by the values of 1.378, 1.178, and 2.106, respectively. These high correlations indicate that the corresponding traits are highly intercorrelated and are not distinctive from each other. These high correlations might absorb a large portion of measurement error for Model I and result in a small chi-square value and non-significant probability. Therefore, a

---

[1] Theoretically, correlations of such magnitude are impossible. These values may be an artifact of "overfitting" too many variables to the models and suggest the need for a new model with fewer factors.

TABLE 9

LISREL Estimates for Discourse Mode and Traits Effects

Model I

	GI	FO	OR	SU	ME	Narrative	EX
GIN	.629 (.109	0	0	0	0	.592 (.130)	0
GIE	.629 (.109)	0	0	0	0	0	.518 (1.33)
F N	0	.580 (.115)	0	0	0	.493 (.133)	0
F E	0	.580 (.115)	0	0	0	0	.508 (.146)
O N	0	0	.194 (.321)	0	0	.822 (.130)	0
O E	0	0	.194 (.321)	0	0	0	.685 (.167)
S N	0	0	0	.374 (.170)	0	.656 (.141)	0
S E	0	0	0	.374 (.170)	0	0	.557 (.157)
M N	0	0	0	0	.726 (.099)	.243 (.124)	0
M E	0	0	0	0	.726 (.099)	0	.433 (.130)

	GI	FO	OR	SU	ME
GI	1.0				
FO	.834 (.140)	1.0			
OR	1.378 (1.814)	2.106 (2.986)	1.0		
SU	1.178 (.341	.552 (.321)	.980 (1.140	1.0	
ME	0.773 (.125)	.429 (.190)	.663 (.963)	.693 (.297)	1.0

χ^2 20 = 19.94
p = 0.4623

refined model is required to fit the structure of the data.

A refined model with reasonable psi-coefficients requires some modifications on the CSE subscales. Signaled by information from psi correlations, we referred to the definitions of the CSE subscales (traits). First, the trait of General Impression is a global rating encompassing, but not limited to the other subscales. The psi coefficients of General Impression with other traits ranging from .773 to 1.378 suggest that GI is highly correlated with each of the four specific traits, supporting the global characteristic of GI. Second, the traits of Focus and Organization have common elements in their definitions. Both traits are basically dealing with the construct of coherence, the logical relationships among ideas in the essay. Since the GI trait is not so distinctive from the others and contributes little information beyond the four specific traits, it seemed advisable to exclude the GI trait from the model. Model I also supports the advisability of collapsing the traits of Focus and Organization into a single trait of Coherence.

Figure 2 presents the refined model. Instead of five traits, the new model consists of only three traits: Coherence, which combines Focus and Organization, and the traits of Support and Mechanics. The LISREL estimates for Model II, summarized in Table 10, show that the value of chi-square, 16.028, again yields a non-significant probability of .248. With

- - - - - - - - - - - -

Insert Figure 2 here

- - - - - - - - - - - -

- - - - - - - - - - - -

Insert Table 10 here

- - - - - - - - - - - -

FIGURE 2

Discourse Mode

Path Diagram for Model II

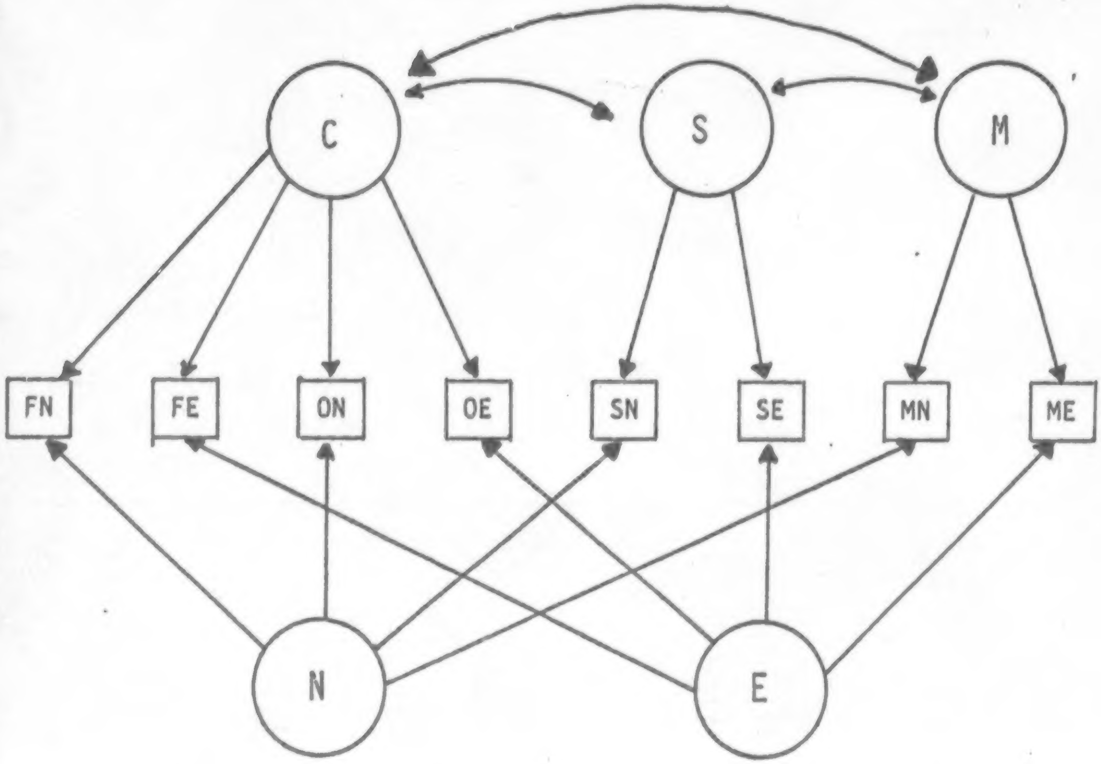

C
S
M
FN
FE
ON
OE
SN
SE
MN
ME
N
E

TABLE 10

LISREL Estimates for Discourse Mode and Refined Trait Effects

	C	S	M	N	E
FN	.559 (.119)	0	0	.595 (.131)	.000
FE	.559 (.119)	0	0	.000	.580 (.146)
ON	.348 (.139)	0	0	.789 (.125)	.000
OE	.348 (.139)	0	0	.000	.599 (.155)
SN	0	.294 (.195)	0	.739 (.145)	.000
SE	0	.294 (.195)	0	.000	.667 (.162)
MN	0	0 (.097)	.742	.285 (.122)	.000
ME	0	0	.742 (.097)	.000	.320 (.133)

C	1.000		
S	.311 (.462)	1.000	
M	.413 (.197)	.854 (.476)	1.000

$\chi^2_{13} = 16.028$

$\rho = 0.248$

Table 11

Discourse Mode

Correlation Matrix for Model III Pooling Essay and Multiple Choice Test Scores

	FON	FOE	ORN	ORE	SUN	SUE
FON	1.0					
FOE	.647	1.0				
ORN	.661	.455	1.0			
ORE	.567	.622	.590	1.0		
SUN	.518	.488	.454	.416	1.0	
SUE	.504	.650	.409	.568	.615	1.0

a well-fitted model, we can proceed to inspect each parameter estimate. The loadings of observed data on corresponding traits are significant, except for the Support subscales. The psi coefficients which are less than 1 and comparatively low, with the respect to the size of their standard errors, imply that the traits of Coherence, Support, and Mechanics, are distinctive from each other. The high loadings of observed data on the two method factors, discourse mode, indicate that the influence of discourse mode is quite large on observed data. This result further supports previous analyses suggesting that the discourse mode required by an essay task might dominate a student's performance.

To examine whether the discourse mode effect holds across response mode as well, a second set of analyses investigated the factor structure underlying subskill (trait) and discourse mode in pooled essay and multiple choice test scores. Six scores were formed for these analyses, two each (expository and narrative) for the three traits common to essay and multiple choice test scores, Focus, Organization and Support. This analysis did not combine Focus and Organization into a Coherence factor since this procedure would have produced a two trait, two-method model not appropriate for the LISREL calculations. The correlation matrix for these variables appears in Table 11.

- - - - - - - - - - -

Insert Table 11 here

- - - - - - - - - - -

Figure 3 displays the path diagram for Model III and Table 12 presents the LISREL estimates of the free and constrained parameters.

- - - - - - - - - - - -

Insert Figure 3 here

- - - - - - - - - - - -

- - - - - - - - - - - -

Insert Table 12 here

- - - - - - - - - - - -

The non-significant probability of .6587 yielded by chi-square of 1.6030 with 3 df., suggests that data pooled across essay and multiple choice scores are adequately explained by this model (Figure 3). In other words, the variances of each of the six observed scores in this model are accounted for by two sources of influence, discourse mode and CSE subscale trait. High loadings of the pooled scores on their corresponding traits (Focus, Organization, or Support) reveal that each score is well defined by its underlying trait. Loadings on two discourse modes (narrative or expository) are low to moderate except for ORN. Apparently, ORN was more affected by discourse mode. In comparison to discourse mode, CSE traits have more dominant effects on pooled scores. The relations among trait factors show that Focus, Organization and Support are moderately to highly correlated, (.697 to .831), supporting the hypothesis that the traits are distinctive, but not independent from one another. The traits Focus and Organization are again highly correlated (.831). In comparison to the psi matrix for Model I, the addition of the multiple choice scores seems to increase the interrelationships among the subscales, suggesting that the multiple choice data blurs the distinctiveness of trait information yielded only by essay ratings. We examined this possibility directly through analyses of response mode effects.

FIGURE 3

Discourse Mode
Path Diagram for Model III

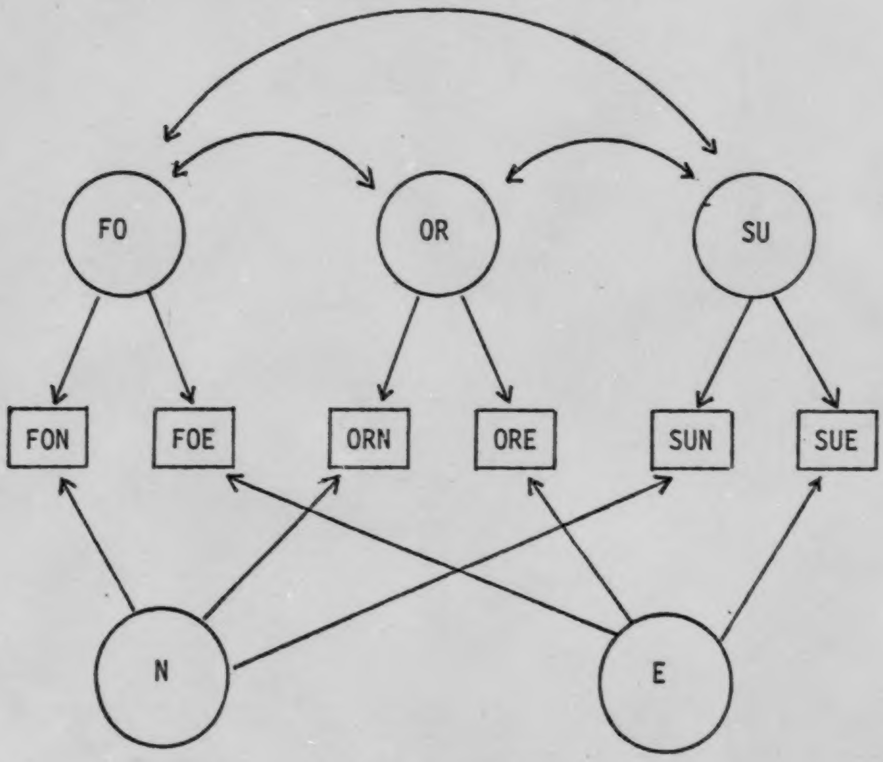

FO
OR
SU
FON
FOE
ORN
ORE
SUN
SUE
N
E

Table 12

LISREL Estimates for Discourse Mode and Refine Trait Effects Pooling Essay and Multiple Choice Scores

Model III

	FO	OR	SU	N	E
FON	.811 (.074)	0	0	.233 (.381)	0
FOE	.811 (.074)	0	0	0	.348 (.135)
ORN	0	.764 (.076)	0	.660 (.980)	0
ORE	0	.764 (.076)	0	0	.300 (.128)
SUN	0	0	.786 (.075)	.080 (1.75)	0
SUE	0	0	.786 (.075)	0	.460 (.156)

	FO	OR	SU	N	E
FO	1.0				
OR	.831 (.062	1.0			
SU	.780 (.070)	.697 (.086)	1.0		

$\chi^2_3 = 1.6030$

$\rho = 0.6587$

Response Mode Effects

The second measurement issue addressed by the study was whether tasks requiring different response modes (direct, production modes: essays, paragraphs; indirect, selection modes: multiple choice) provide the same type and quality of data about student writing abilities. Analyses first examined correlations between scores students received on their essays, paragraphs and multiple choice questions when the scores were from measures all in the same discourse mode and when the scores were on tests in different discourse modes. Table 13 presents these comparisions.

- - - - - - - - - - - -

Insert Table 13 here

- - - - - - - - - - - -

In general the relationship between scores a student received on the two direct measures (essays and paragraphs) are stronger than the relationship between his/her scores on the direct and indirect measures. Furthermore, the substantially lower correlations between response mode scores when the tasks also differ in discourse mode further corroborates the discourse mode hypotheses. An MTMM analyses was performed next to search more intensively for the factor structure underlying response mode effects. The method variables for response mode were constructed according to procedures analogous to those used to construct the discourse mode variables. Each of the five subscale scores provided by the essay and paragraph ratings was averaged over raters, then standardized within topic and discourse mode. Standardizations were accomplished by removing possible interactions between response mode on one hand and genre and/or topic on

Table 13

Correlations Between Test Scores from Different Response Modes

All Tests Were in the Same Discourse Mode

	EE	EP	EMc	PMc
GI	.581 (78)	.147 (153)	X	X
FO	.559 (78)	.275 (153)	.314 (316)	.371 (192)
OR	.473 (78)	.208 (153)	.152 (316)	.300 (192)
SU	.286 (78)	.353 (153)	.136 (316)	.323 (192)
ME	.629 (78)	.582 (153)	X	X
Total	.624 (78)	.428 (153)	.326 (316)	.471 (192)

Tests Were in Different Response Modes

	EE	EP	EMc	PMc
GI	.126 (94)	.415 (90)	X	X
FO	.391 (94)	.205 (90)	.211 (316)	.288 (192)
OR	.008 (94)	.301 (90)	.180 (316)	.233 (192)
SU	.229 (94)	.345 (90)	.180 (316)	.203 (192)
ME	.583 (94)	.632 (90)	X	X
Total	.355 (94)	.484 (90)	.325 (316)	.376 (192)

GI - General Impression
FO - Focus
OR - Organization
SU - Support
ME - Mechanics

EE - Essay: Essay
EP - Essay: Paragraph
EMc- Essay: Multiple Choice
PMc- Paragraph: Multiple Choice

the other. Complete data were available for 148 of the students.

Number correct scores were formed within genre for each of the three multiple choice subscales, standardized within genre, summed across genre, and then restandardized to produce scores scaled in a manner comparable to those derived from the writing samples. No measures of General Impression and Mechanics were included in the multiple choice task.

In sum, 18 scores were constructed for analysis, three measures each of General Impression and Mechanics (two essay and one paragraph), and four each of Focus, Organization and Support (two essay, one paragraph and one multiple choice). The variables and the abbreviations used to refer to them in the response mode analyses appear in Table 14.

- - - - - - - - - - - -

Insert Table 14 here

- - - - - - - - - - - -

The transformed scores permitted analysis first of the relationships among the response mode variables after removing genre effects. Table 15 presents the correlations matrix for the three response modes and five subscales.

- - - - - - - - - - - -

Insert Table 15 here

- - - - - - - - - - - -

The MTMM analyses began by considering the data for the "Essay 1" and "Essay 2" methods only, examining the eight scores defined for these two conditions, two measures of Focus (fe_1, fe_2), Organization (oe_1, oe_2), Support (se_1, se_2), and Mechanics (me_1, me_2). While the models consider Coherence as the combination of Foucs and Organization, their separate observed scores are entered into the analysis. The model specified for

Table 14

Description of Variables in LISREL Analyses of Response Mode Effects

WRITING VARIABLES	"Essay 1"	"Essay 2"	Paragraph	Multiple Choice
General Impression	gie_1	gie_2	gip	---
Focus	fe_1	fe_2	fp	fmc
Organization	oe_1	oe_2	op	omc
Support	se_1	se_2	sp	smc
Mechanics	me_1	me_2	mp	---

Table 15

Correlation Matrix of Scores Containing Response Mode and Subscale Trait Effects

	GIE_1	GIE_2	GIP	FOE_1	FOE_2	FOP	FOMC	ORE_1	ORE_2	ORP	ORMC	SUE_1	SUE_2	SUP	SUMC	MEE_1	MEE_2	MEP
GIE_1	1.000																	
GIE_2	0.282	1.000																
GIP	0.333	0.232	1.000															
FOE_1	0.407	0.318	0.236	1.000														
FOE_2	0.215	0.521	0.271	0.424	1.000													
FOP	0.132	0.180	0.511	0.231	0.270	1.000												
FOMC	0.214	0.325	0.299	0.256	0.379	0.362	1.000											
ORE_1	0.816	0.197	0.224	0.441	0.211	0.057	0.226	1.000										
ORE_2	0.281	0.775	0.250	0.345	0.608	0.192	0.353	0.252	1.000									
ORP	0.235	0.209	0.810	0.195	0.279	0.559	0.286	0.176	0.228	1.000								
ORMC	0.131	0.285	0.348	0.229	0.360	0.304	0.474	0.142	0.307	0.339	1.000							
SUE_1	0.670	0.274	0.253	0.434	0.247	0.252	0.149	0.566	0.230	0.223	0.181	1.000						
SUE_2	0.284	0.525	0.179	0.233	0.475	0.186	0.215	0.246	0.550	0.170	0.256	0.314	1.000					
SUP	0.268	0.360	0.620	0.324	0.375	0.520	0.344	0.269	0.322	0.579	0.328	0.353	0.294	1.000				
SUMC	0.314	0.165	0.302	0.239	0.264	0.259	0.419	0.312	0.253	0.291	0.437	0.197	0.172	0.245	1.000			
MEE_1	0.492	0.286	0.305	0.423	0.234	0.305	0.222	0.436	0.264	0.232	0.285	0.396	0.266	0.297	0.317	1.000		
MEE_2	0.357	0.452	0.392	0.365	0.512	0.353	0.373	0.276	0.419	0.297	0.344	0.300	0.361	0.428	0.354	0.589	1.000	
MEP	0.369	0.298	0.478	0.345	0.352	0.433	0.311	0.311	0.399	0.429	0.352	0.314	0.272	0.330	0.360	0.603	0.555	1.000

these variables includes the three relatively distinct trait/subscale content factors emerging from the discourse mode analyses and two "method" factors, one for each essay. Figure 4 displays a path diagram for Model I, and Table 16 presents the LISREL estimates of the free and constrained parameters for the model.

- - - - - - - - - - - - - - - - -

Insert Figure 4 and Table 16 here

- - - - - - - - - - - - - - - - -

Similar to the model specified for the discourse mode MTMM analysis, the Figure shows Model I allowing the trait or subscale factors to be freely intercorrelated, while the method factors are specified to be uncorrelated with each other and with the subscale factors. The restriction on the method factor correlations corresponds to the hypothesis that they act as independent additive components in the explanation of the observed scores. In addition, we constrained the factor loadings for each pair of subscale measures on their corresponding trait factors to equal one another. These constraints are equivalent to a test of the hypothesis that subscale scores from different essays will exhibit the same degree of relationship to the trait factor they measure.

The model as a whole cannot be rejected; the chi square goodness of fit test yields a probability of .202 (ns), suggesting that the model provides an adequate account for the observed correlations among essay variables. Loadings of the essay variables on their respective trait factors are all substantial and highly significant, ranging from a low of .472 for Organization to a high of .768 for Mechanics. The Organization subscale

FIGURE 4

Response Mode

Path Diagram for Model I

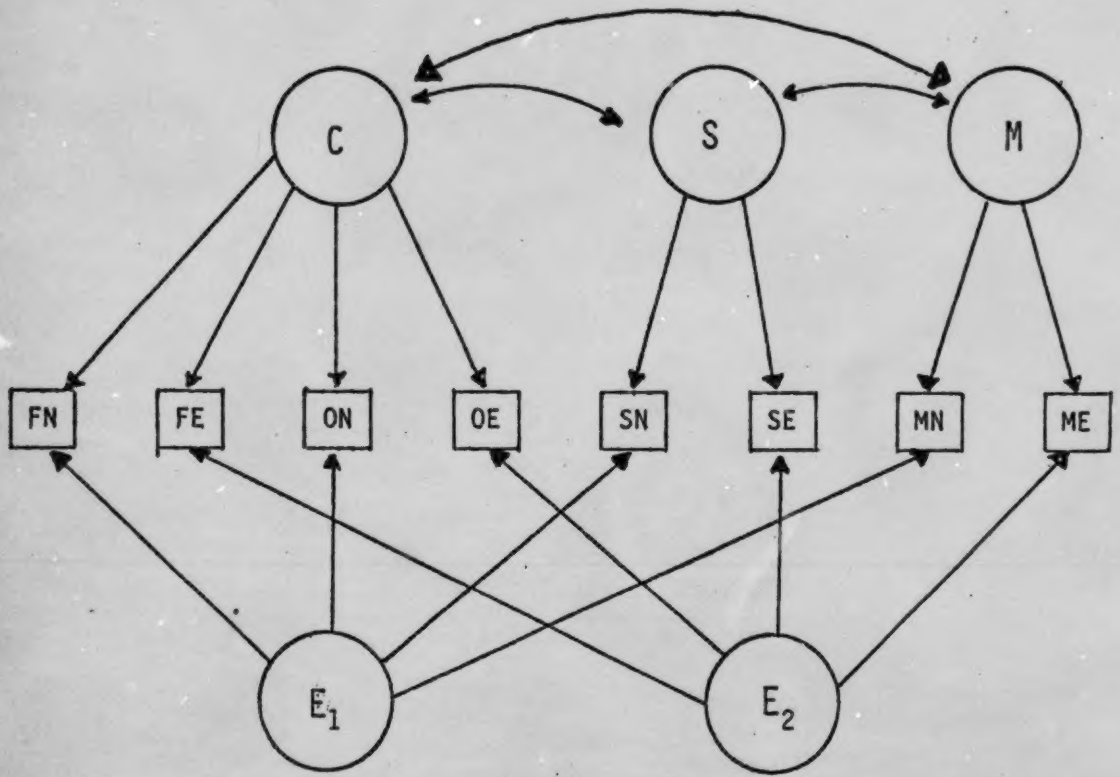

C
S
M
FN
FE
ON
OE
SN
SE
MN
ME
E_1
E_2

Table 16

LISREL Estimates for Response Mode and Trait Effects

Model I

LAMBDA	C	S	M	E_1	E_2
Fe_1	.633 (.069)	0	0	.255 (.100)	0
Fe_2	.633 (.069)	0	0	0	.485 (.095)
Oe_1	.472 (.075)	0	0	.671 (.130)	0
Oe_2	.472 (.075)	0	0	0	.639 (.105)
Se_1	0	.535 (.077)	0	.534 (.118)	0
Se_2	0	.535 (.077)	0	0	.507 (.104)
Me_1	0	0	.768 (.062)	.289 (.092)	0
Me_2	0	0	.768 (.062)	0	.254 (.089)

PSI	C	S	M
C	1		
S	.799 (.102)	1	
M	.706 (.083)	.626 (.115)	1

χ^2 with df 13 = 16.9398

ρ = .2021

loadings on the method factors are relatively high and the loadings of the Support subscale are moderate. Both Focus and Mechanics loadings are low. Turning to the psi matrix, we see that the estimates of the relations among the trait factors are moderate (below .80), ranging from a low of .626 for the correlation between Mechanics and Support to a high of .799 between Coherence and Support. The Mechanics factors appears to be the most independent in the set.

Model II adds the data from the paragraph task and expands to 12 the number of variables included in the analysis. Four subscales forming the three trait factors specified in Model I each have, under Model II, an additional subscale score loading (no constraints are placed on these loadings); and a new, second method factor appears, Paragraph, to account for covariation specific to this mode of responding. Figure 5 displays the path diagram for this model and Table 17 presents the results of the LISREL estimation of the parameters of Model II.

- - - - - - - - - - - -

Insert Figure 5 here

- - - - - - - - - - - -

- - - - - - - - - - - -

Insert Table 17 here

- - - - - - - - - - - -

Model II provides an adequate overall fit to the observed intercorrelations among the essay and paragraph variables (chi square with 43 df = 51.533, $\underline{p}$ = .175). This result provisionally supports the hypothesis that the scores generated by application of the rating system to paragraph length writing samples can be interpreted as measuring the same underlying

Figure 5

Response Mode

Path Diagram for Model II

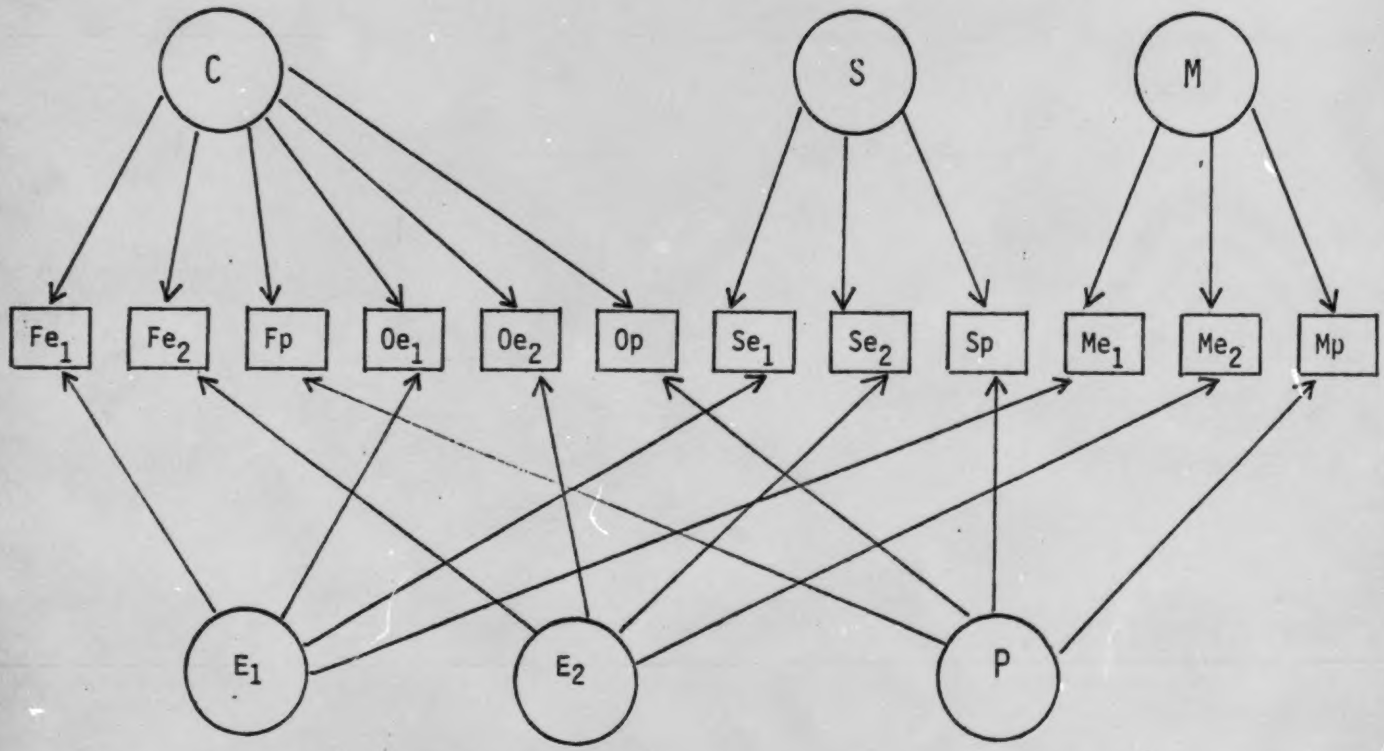

C
S
M
Fe$_1$
Fe$_2$
Fp
Oe$_1$
Oe$_2$
Op
Se$_1$
Se$_2$
Sp
Me$_1$
Me$_2$
Mp
E$_1$
E$_2$
P

Table 17

LISREL Estimates for Response Mode and Trait Effects

Model II

LAMBDA	C	S	M	E_1	E_2	P
Fe_1	.581 (.068)	0	0	.282 (.096)	0	0
Fe_2	.581 (.068)	0	0	0	.495 (.094)	0
Fp	.485 (.094)	0	0	0	0	.473 (.100)
Oe_1	.468 (.069)	0	0	.683 (.126)	0	0
Oe_2	.468 (.069)	0	0	0	.650 (.104)	0
Op	.417 (.095)	0	0	0	0	.780 (.114)
Se_1	0	.522 (.070)	0	.502 (.111)	0	0
Se_2	0	.522 (.070)	0	0	.461 (.099)	0
Sp	0	.639 (.095)	0	0	0	.421 (.095)
Me_1	0	0	.773 (.062)	.250 (.081)	0	0
Me_2	0	0	.773 (.062)	0	.192 (.080)	0
Mp	0	0	.728 (.078)	0	0	.207 (.078)

PSI	C	S	M
C	1		
D	.937 (.061)	1	
M	.809 (.066)	.684 (.082)	1

χ^2 with df of 43 = 51.5331

ρ = .175

content as the scores derived from full length essays. Inspection of the lambda matrix shows that the loadings for paragraph subscale scores on their associated trait factors are of substantial magnitude in each case, and that the loadings on the paragraph factor follow the same general pattern as for the two essay method factors. With one exception, the paragraph variables appear to relate to trait factors less strongly than do the essay scores. The one exception is an interesting one: "sp" provides a clearer definition of the Support factor than either of the support measures derived from essays. This would seem to suggest that the rater's task of judging the use of support is carried out more distinctly in the context of single paragraphs than it is in longer writing samples. A test of this hypothesis, however, would require multiple paragraph measures of the "sp" variable.

As in Model I, the trait intercorrelations in the Model II Psi matrix are moderate to high, indicating considerable interdependence among the subscales. Again, Mechanics exhibits lower levels of relationship to the other subscales.

Comparison of Models I and II reveals two main differences. First, there is some instability in the size of the essay variables' loadings on the associated trait factors as we move from the first to the second model. This leads to the interpretation that the factors composed of both essay and paragraph variables do not measure precisely the same content as factors composed of essay variables only. (In order to retain the same factor traits across both models, we can constrain the lambda coefficients and psi coefficients in Model I to be identical to corresponding coefficients in Model II. This procedure is applied in later models.) Second, estimates of the trait intercorrelations in Model II are greater than their counterparts in Model I. Thus, although the inclusion of paragraph scores may have broadened the content of the trait factors, it seems also to have diminished their distinctiveness.

The third MTMM analysis builds on the previous two by adding the three scores derived from the multiple choice items administered to the students in the study. Recall that only items analogous to the Focus, Organization and Support subscales were included in the multiple choice test. Model III differs from Model II, then, by the specification of trait loadings for these three subscales, and the addition of a multiple choice method factor. Figure 6 displays the path diagram for Model II and Table 18, the LISREL estimates of the model parameters.

- - - - - - - - - - - - - - - - - -

Insert Figure 6 and Table 18 here

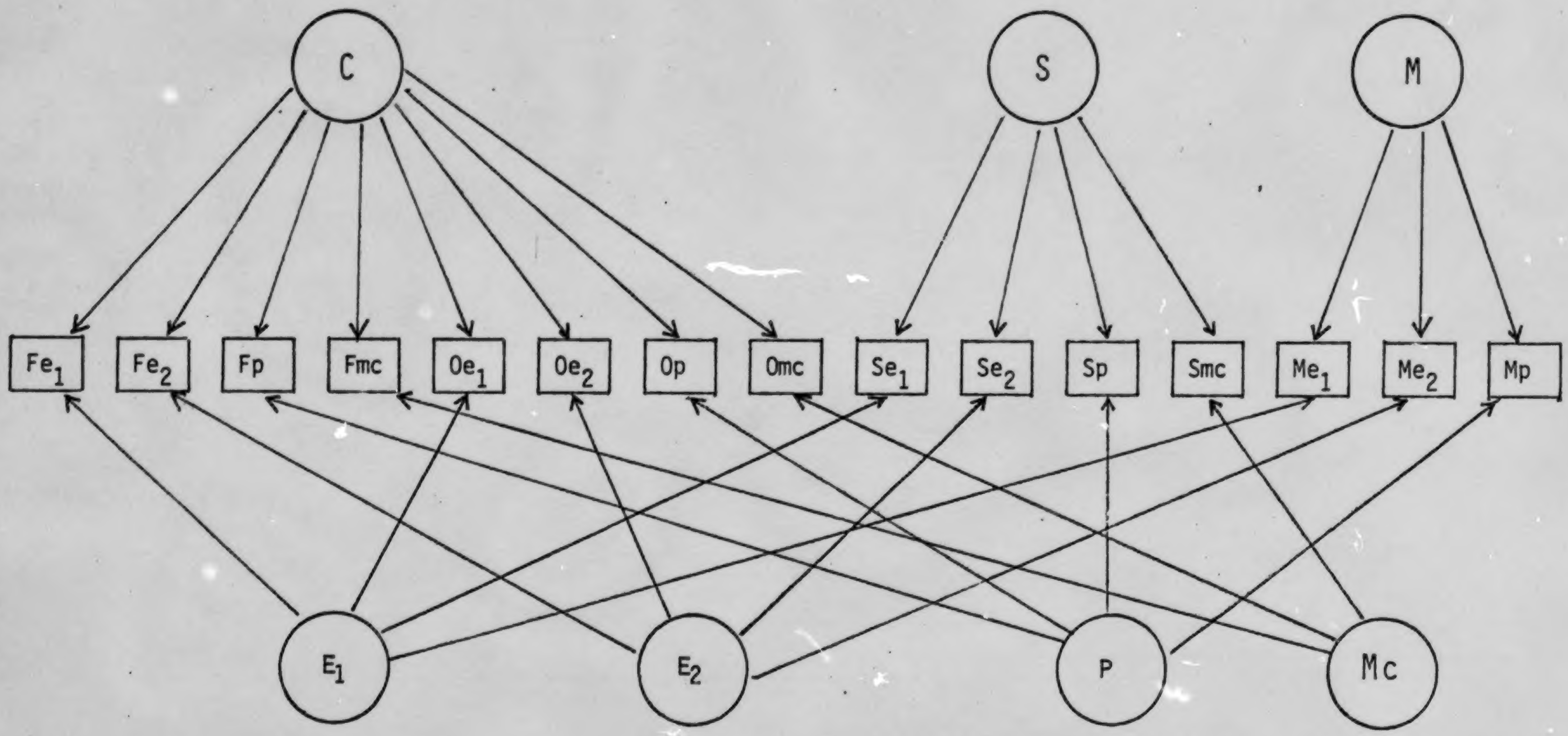
- - - - - - - - - - - - - - - - - -

As in the first two analyses, Model III provides a reasonably good fit to the data (chi square with 76 df = 84.962, $\underline{p}$ = .226), implying that the same 3-trait structure is not violated by the inclusion of the multiple choice scores. By and large, however, the Model III estimates of the essay trait factor loadings have dropped in value in comparison with the corresponding estimates from Models I and II. Also, the trait factor intercorrelations in Model III have increased for Coherence and Mechanics, indicating that the subscale content factors has drifted closer together as a result of adding the multiple choice variables. Thus, while the multiple choice scores apparently share some content with the constructed response variables to which they are purportedly analogous, they also seem to possess a higher degree of "latent collinearity" (Yates, 1979) in the trait factor

Figure 6

Response Mode

Path Diagram for Models III and IV


C
S
M
Fe$_1$
Fe$_2$
Fp
Fmc
Oe$_1$
Oe$_2$
Op
Omc
Se$_1$
Se$_2$
Sp
Smc
Me$_1$
Me$_2$
Mp
E$_1$
E$_2$
P
Mc

27a

Table 18

LISREL Estimates for Response Mode and Trait Effects

Model III

	C	S	M	E_1	E_2	P	Me
Fe_1	.585 (.065)	0	0	.323 (.092)	0	0	0
Fe_2	.585 (.065)	0	0	0	.472 (.090)	0	0
Fp	.514 (.089)	0	0	0	0	.444 (.100)	0
Fmc	.546 (.088)	0	0	0	0	0	.399 (.125)
Oe_1	.483 (.066)	0	0	.672 (.111)	0	0	0
Oe_2	.483 (.066)	0	0	0	.619 (.100)	0	0
Op	.470 (.090)	0	0	0	0	.737 (.118)	0
Omc	.528 (.089)	0	0	0	0	0	.462 (.136)
Se_1	0	.469 (.067)	0	.546 (.103)	0	0	0
Se_2	0	.469 (.067)	0	0	.480 (.098)	0	0
Sp	0	.611 (.089)	0	0	0	.398 (.094)	0
Smc	0	.483 (.091)	0	0	0	0	.397 (.128)
Me_1	0	0	.772 (.061)	.277 (.078)	0	0	0
Me_2	0	0	.772 (.061)	0	.183 (.078)	0	0
Mp	0	0	.747 (.077)	0	0	.182 (.078)	0

	C	S	M
C	1		
S	.978 (.049)	1	
M	.786 (.059)	.786 (.073)	1

X^2 with 76 df = 84.9619

P = 0.2255

space. Whether this situation arises because the multiple choice variables are related to writing ability in some non-specific fashion, or because all of the variables, but especially the multiple choice scores, share a common dependence on general verbal ability, can not be disentangled without additional analyses, including tests marking general ability factors. In any event, it is reasonable to interpret the increased interdependence among trait factors as an indication that the multiple choice scores possess generally lower validity as indices of distinctive components of writing ability than do measures based on actual writing samples.

Model IV examines the relationship of the paragraph and multiple choice variables to the set of trait factors defined solely on the basis of the essay variables. Generally speaking, both the contribution of the essay scores to the definition of the subscale content factors and the degree of independence of these factors from one another were reduced as data from the alternative response modes were added to the analysis. Model IV takes the trait factor structure that obtained when only essay scores were included in the analysis (i.e., Model I) as a criterion definition of the content underlying the subscales, treating the Model I trait factors as "unmeasured" criterion variables against which to compare the scores from the other two response modes. This can be accomplished in LISREL by modifying the specification for Model II in two ways. First, instead of estimating trait loadings for essay variables, new specifications fix their values to equal those estimated in Model I. Second, we place a similar constraint on the trait factor intercorrelations in Psi, by fixing their values at those obtained in the essay-only solution for Model I. These two sets of restrictions will ensure that the trait factors found in Model I will be reproduced exactly in Model IV, and the standing of the paragraph and multiple choice variables can be evaluated vis-à-vis the essay criterion trait structure. The LISREL estimates of the free parameters in Model IV are contained in Table 19.

Insert Table 19 here

The only parameter estimates of direct interest in Table 19 are the trait factor loadings for the paragraph and multiple choice variables. The data indicate near uniform reduction in their magnitude in comparison to the estimates obtained from Model III. This shift does not reduce the overall model fit (chi square with 83 df = 95.547, $\underline{p}$ = .164). In all but two instances, paragraph and multiple choice trait factor loadings are lower than the corresponsing loadings for the essay variables. Both exceptions are recurrences of the findings from Model II and Model III that the measure of Support derived from a paragraph length writing sample outperforms the Support measures based on full length essays and the Organization score in multiple choice is slightly more distinct than the Organization measures on essays respectively. Support as measured by multiple choice items seems to reflect relatively little of what is measured in actual writing samples; while multiple choice measures of Focus and Organization seem to convey a roughly comparable amount of information about subscale content to that contained in a single paragraph.

Table 19

LISREL Estimates for Response Mode and Trait Effects

Model IV

	C	S	M	E_1	E_2	P	Mc
Fe_1	.633	0	0	.333 (.092)	0	0	0
Fe_2	.633	0	0	0	.436 (.088)	0	0
Fp	.463 (.088)	0	0	0	0	.457 (.095)	0
Fmc	.526 (.087)	0	0	0	0	0	.407 (.115)
Oe_1	.472	0	0	.676 (.109)	0	0	0
Oe_2	.472	0	0	0	.606 (.100)	0	0
Op	.408 (.088)	0	0	0	0	.767 (.113)	0
Omc	.492 (.088)	0	0	0	0	0	.452 (.123)
Se_1	0	.535	0	.538 (.099)	0	0	0
Se_2	0	.535	0	0	.496 (.100)	0	0
Sp	0	.604 (.089)	0	0	0	.414 (.089)	0
Smc	0	.410 (.094)	0	0	0	0	.475 (.131)
Me_1	0	0	.768	.290 (.078)	0	0	0
Me_2	0	0	.768	0	.178 (.079)	0	0
Mp	0	0	.719 (.072)	0	0	.197 (.075)	0

χ^2 with 83 df = 95.5468

$\rho = 0.1636$

Summary and Conclusions

The purpose of the study was to examine the comparability of writing competency profiles derived from test tasks differing in discourse and response mode. Theory and research in the fields of learning, instruction and rhetoric have fueled contentions that the knowledge structures and processing strategies activated by different writing aims and modes of responding are quite distinct. We were attempting to demonstrate the robustness of these claims from a measurement perspective.

In practice, many current writing assessment programs fail to consider the validity of test data that does not distinguish between the demands of types of writing tasks and between the requirements of production and selection. At heart, the issue is one of construct validity, do these alternative task and processing variables measure the same thing? Our results indicate that the answer is "no."

In this study the results of correlational, parametric and multitrait multimethod analyses indicate that levels of performance vary on tasks presenting different writing purposes. These data cast doubt on the assumption that "a good writer is a good writer" regardless of the assignment. The implication is that writing for different aims draws different skill constructs which must therefore be measured separately to avoid erroneous, invalid interpretations of performance. The findings suggest that generalizations about student writing competence must reference the particular discourse domain rather than the general domain of writing.

The study also investigated the distinctiveness of information about writing competence provided by direct and indirect measurement techniques. Again the issue is one of validity, do both response modes measure the same skill construct? Again the answer is "no." When essay performance is set as the criterion, multiple choice performance seems to be a poor proxy. Tests of response mode effects within an MTMM framework suggest that scores on the General Impression and Organization subscales contain large method components when measures are taken from constructed responses. A plausible explanation for this finding is that the method factor loadings for these variables are inflated by within-occasion (e.g., a given essay) residual linkages between GI and O brought about by raters' tendency to depend more on Organization than on other specific features in formulating their General Impression rating. The remaining three subscales all were found to contain proportionately larger amounts of content-related variance than method related variance, with Mechanics appearing to be the purest of the three. The patterning of method variance saturation in the five subscales was the same for the three writing samples available for each subject.

An interesting picture of the effects of the varying response mode emerged from the analyses. While models can be fitted to the data from all three response modes that confirm the subscales' content, the degree of independence of the resulting subscale factors appears to be affected by which response modes are included in the analysis. The most differentiated subscale factor structure is obtained by including only essay variables in the analysis; interdependence among the subscale factors increases with the addition of both paragraph and multiple choice measures. Thus, the effect of shortening the assessment task for the examinee through examination of just paragraphs or of changing the form of the response (multiple choice tasks) does not simply increase the measurement error. The savings in testing time are obtained also at the cost of clarity and distinctiveness in the information about each of the subscales. When the subscale content factors are located in the variable space so as to maximize their relationship to scores derived from the essay response mode, all other subscale-response mode combinations, except one, provide weaker substantive information. The one exception is the measure of Support based on paragraph-length writing samples which seems to be superior to the corresponding essay variables in its ability to capture subscale content. It may be that the use of support is less equivocally evaluated in the context of a single paragraph than in an essay containing multiple paragraphs, each of which may suggest a different view of the examinee's ability to provide supporting detail.

The MTMM analyses also provided information about the validity of the rating scales. The MTMM analyses suggested, first, that repeated applications of the CSE scoring rubric to writing samples in fact produce measures that tap the same underlying content. Thus, given multiple measures of each subscale, it is possible to fit a factor analysis model that confirms their hypothesized content. Second, it was found that factors reflecting the content of all five subscales are strongly intercorrelated, and this interdependence appears to be present no matter what response mode subjects are assessed in. When the global judgment for General Impression was removed and Focus and Organization combined into a Coherence subscale, scale intercorrelations became more moderate and distinct. Since techniques for producing writing that is coherent, supported and mechanically correct are often taught separately, further examination of the value of rating writing according to separate component features should consider both their diagnostic utility and component distinctiveness.

Finally, the study exemplified the contribution MTMM analyses can make to validity studies. The technique may provide more sensitive, precise statistical indices of hypothesized competencies underlying test performance.

In summary, the study highlights the importance of precision in designing, analyzing and reporting writing assessment data. It may be that the techniques developed for specifying domain-referenced skill boundaries can provide a reasonable framework for focusing attention, discussion, assessment and instruction on clearly bounded classes of writing performance.

References

Anderson, R. C. The notion of schemata and the educational enterprise. In Anderson, R. C., Spiro, R. J., & Montague, W. E. (Eds.), Schooling and the acquisition of knowledge. Hillsdale, NJ: Lawrence Earlbaum Associates, 1977.

Bourne, L. J. Human conceptual behavior. Boston: Allyn and Bacon, 1966.

Braddock, R., Lloyd-Jones, R., & Schoer, L. Research in written composition. Champaign, IL: National Council of Teachers of English, 1963.

Breland, H. M. A study of college English placement and the test of standard written English. Princeton, NJ: Educational Testing Service, January, 1977.

Breland, H. M., Conlan, G. C., & Ragosa, D. A preliminary study of the Test of Standard Written English. Princeton, NJ: Educational Testing Service, 1976.

Breland, H. M., & Gaynor, J. L. A comparison of direct and indirect assessments of writing skill. Journal of Educational Measurement, 1979, 16, 119-128.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validity by the multitrait-multimethod matrix. Psychological Bulletin, 1979, 56, 81-105.

Coffman, W. E. Essay exams. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D. C: American Council on Education, 1971.

Cooper, C. R. Current studies of writing achievement and writing competence. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Cooper, C., & Odell, L. (Eds.) Evaluating writing: Describing, measuring, judging. Buffalo, NY: State University of New York at Buffalo, 1977.

Crowhurst, M. Syntactic complexity in narration and argument at three grade levels. Canadian Journal of Education, 1980.

Diederich, P. B. Measuring growth in English. Champaign, IL: National Council of Teachers of English, 1974.

French, J. W. Schools of thought in judging excellence of English themes. 1961 Proceedings of Invitational Conference on Testing Problems, Educational Testing Service, Princeton, NJ, 1962.

Godshalk, F. I., Swineford, F., & Coffman, W. E. The measurement of writing ability. New York: College Entrance Examination Board, 1966.

Graesser, A. C., Hauft-Smith, K., Cohen, A. D., & Pyles, D. Familiarity and test genre on retention of prose. Journal of Experimental Education, 1979, 48, 281-290.

Hogan, T. P., & Mishler, C. Relationships between essay tests and objective tests of language skills for elementary school students. Journal of Educational Measurement, 1980, 17, 219-227.

Jöreskog, K. G. Analyzing psychological data by structural analysis of covariance matrices. In R. C. Atkinson, D. H. Krantz, & P. D. Suppes (Eds.), Contemporary Development in Mathematical Psychology, Vol. II. San Francisco: W. H. Freeman & Co., 1974, 1-56.

Kinneavy, J. L. A theory of discourse. The Aims of Discourse. Englewood Cliffs, NJ: Prentice Hall, Inc., 1971.

Meyer, B. F. The organization of prose and its effects on memory. North Holland Studies in Theoretical Poetics (Vol. I). Amsterdam: North Holland Publishing Company, 1975.

Minsky, M. A framework for representing knowledge. In P. H. Winston (Ed.), The psychology of computer vision. New York: McGraw-Hill, 1975.

Perron, J. D. Written syntactic complexity in modes of discourse. Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

Pitts, M. The relationship of classroom instructional characteristics and writing in the descriptive/narrative mode. Report to the National Institute of Education, November, 1978. (Grant No. OB-NIE-G-78-0213 to the UCLA Center for the Study of Evaluation.)

Praeter, D., & Padia, W. Effects of modes of discourse in writing performance in grades four and six. Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

Quellmalz, E. Interim Report. Defining writing domains: Effects of discourse and response mode. Center for the Study of Evaluation, 1979.

San Jose, C. P. M. Grammatical structures in four modes of writing at fourth grade level. Unpublished doctoral dissertation, Syracuse University, 1972. Dissertation Abstracts International, 1973, 33, 5411-A.

Skinner, B. F. Teaching machines and programmed learning. New York: Appleton Century Crofts, 1963.

Smith, L. S. Investigation of writing assessment strategies. Report to the National Institute of Education, November, 1978. (Grant No. OB-NIE-G-78-0213 to the UCLA Center for the Study of Evaluation.)

Traub, R. E., & Fisher, C. W. On the equivalence of constructed-response and multiple-choice tests. Applied Psychological Measurement, 1977, 1(3), 355-369.

Veal, L. R., & Tillman, M. Mode of discourse variation in the evaluation of children's writing. Research in the teaching of English, 1971, 5, 37-45.

Werts, C. E., Jöreskog, K. G., & Linn, R. L. A multitrait-multimethod model for studying growth. Educational and Psychological Measurement, 1972, 32, 655-678.

Winters, L. The effects of differing response criteria on the assessment of writing competence. Report to the National Institute of Education, November, 1978. (Grant No. OB-NIE-G-78-0213 to the UCLA Center for the Study of Evaluation.)

Yates, A. Distinguishing trait from method variance in fitting the invariant common factor model to observed intercorrelations among personality rating scales. Paper presented at the annual meeting of the Society of Multivariate Experimental Psychology, Los Angeles, CA, 1979.

PROBLEMS IN STABLIZING THE JUDGMENT PROCESS

Edys Quellmalz

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endoresement by the National Institute of Education should be inferred.

Problems in Stablizing the Judgment Process

Edys Quellmalz

Center for the Study of Evaluation
University of California, Los Angeles

The increasing demand for competency assessments of complex human performance has led to renewed scrutiny of the conceptual and technical quality of prevailing testing practice. Particularly in the area of language production, i.e., writing, oral language and oral reading, researchers and practitioners assert that competency tests must provide tasks that match performance objectives and that activate cognitive processing strategies required by production rather than recognition tasks. The validity of indirect (i.e., multiple choice) measures is no longer logically, psychologically or ecologically acceptable to the majority of professionals in writing instruction and evaluation. Life is not a multiple choice. Students' language production skills, in particular, must be sufficiently proficient for students to function autonomously in the real world.

Although collecting samples of complex performance can presumably provide "direct," valid measures of content, the renowned unreliability of judging constructed responses continues to plague assessment methodology. Because direct performance samples are mediated by highly variable judgments of raters who score or characterize performance samples along some dimensions, a critical goal for performance judgment in general, and for writing judgment in particular, is to find ways to assure that judges apply scoring criteria accurately and fairly. As a part of a broader program studying issues in test design, we have investigated dimensions of the test tasks, context and scoring that will reduce irrelevant variability in examinee and rater behavior.

This paper analyzes a series of measurement problems that jeopardize the validity of the judgment process and examines the effectiveness of methods currently employed to address these problems. Reviews of prevailing rating practices, in conjunction with cumulative empirical evidence on factors influencing judgments in domain-referenced assessment, demonstrate that direct writing assessment faces a dual validity requirement. Both the test task and the scoring procedure must meet separate conceptual and statistical validity standards. The paper elaborates the requirements for accurate and fair writing competence assessment and illustrates how state-of-the-art rating processes pose serious threats to the validity of the writing assessments.

Domain-Referenced Scoring Requirements

The avowed intent and structure of competency or domain-referenced tests require explicit, replicable scoring criteria and procedures; thus, the need for methods to stabilize rating criteria and readers' application of them is immediate and real. Soon the uniform application of performance criteria may become a legal requirement when decisions based on these tests result in life-altering consequences for students. Mandates proliferate at state and local levels for writing assessment at all levels of public school, and large numbers of writing samples must be scored by great numbers of raters. Many assessment programs are required to provide students repeated opportunities to pass comparable forms of a test. Also built into many assessment programs is a requirement to administer comparable tests, at regular intervals, at geographically separate sites.

The purpose of these competency assessments is to monitor development of students' skills at points specified throughout their schooling, to detect skills for which they might need remedial assistance and to document skill development. A student who fails to demonstrate competency in writing, receives additional instruction, and is then retested should be judged according to the same standards at each test administration. His or her score should not depend on either the performance of a new cohort of examinees nor upon the idiosyncratic values of differently oriented sets of raters.

Unfortunately, many writing assessment programs derive their guidelines from norm-referenced test methodology. In practice, norm-referenced writing tests are scored by ranking papers within the limits of a particular sample. Essays are usually scored holistically, on generally described criteria, and involve scoring procedures where raters rank essays by sorting them into piles anchored by the range of quality of that particular sample (Conlan, 1976). Thus a particular paper's rank and/or score could change from sample to sample, if the range of the quality of the competition varied from one test group to the next. Such practices result in a "sliding scale" where the rated quality of a particular paper changes according to the quality range of papers in the group. For example, a student might take a writing competency test in the fall, when all students, low achievers to college preparatory students, participate. A student's rank in this wide quality range is below mastery. In the spring, the student, along with the restricted range of students who failed the first administration, takes another writing competency test and just passes. Does s/he pass because intervening writing instruction has strengthened weak writing skills, or because her or his rank is higher in the restricted range of poorer writers? Present holistic scoring procedures can not provide an answer to this question. The holistic score provides no evidence of the developmental level of specific writing weakness that were low and may have improved. Despite the use of "anchor" papers during training to illustrate what a "6" or "3" had been for other groups, the most prevalent holistic scoring procedures still require raters to distribute papers across the score range.

A major measurement problem confronting many competency based writing assessments, then, is the failure to deal with the need to assure comparability of scoring between test occasions as well as within a scoring session. Such comparability would require not just statistical indices of rater agreement but comparisons of mean scores, since ratings within a session might agree but differ between sessions. Adopting a norm-referenced method of criteria application based on ranking within occasion imperils, if not precludes between-occasion uniformity of criteria application. Therefore two measurement problems inhere in judgment stability, stability within a session and stability across sessions.

To document scale stability, an assessment would have to intersperse anchor papers scored in previous assessments among papers rated within an on-going rating session and report comparability of anchor paper scores across test occasions and rater groups. Such documentation of comparaability is conspicuously absent in both research and practice.

Research on Rating Variability

Evidence pointing to the sources and manifestations of scale instability can be found in the rapidly accumulating body of research on issues of rating variability. The instability of ratings has been a major, and generally acknowledged, weakness of measures of writing skill (Coffman, 1971b). Braddock, Lloyd-Jones and Schoer (1963) classified four sources of error: 1) the writer, 2) the assignment, 3) the rater, 4) between raters. Although considerable research within the framework of domain-referenced testing has examined dimensions of the test task that influence writer performance such as discourse aim and topic modality (Pitts, 1978; Spooner-Smith, 1978; Quellmalz, 1979; Praeter & Padia, 1980; Crowhurst, 1980), less attention has been given to the factors involved in rater behavior.

In the broadest sense, inter- and intra-rater variability are a matter of fluctuating standards of judgment. Research has amply demonstrated that anarchical scoring of essays, where raters apply their individual standards, results in high disagreement among raters from different occupations (Diederich, French, & Carlton, 1961) and even among English professors (Findlayson, 1951; McColly, 1970). Follman and Anderson (1967) demonstrated that the more homogeneous the background of raters, the more their scoring agreed. Long ago, Eels (1930) demonstrated the problem of intra-rater criteria bias when he found that the variability in essay scores assigned even by the same reader on different occasions approached the degree of variability of scores assigned by different readers. Recognizing the magnitude of error occurring in unstructured scoring, researchers attempted to devise various techniques for controlling score variability.

Methods for Controlling Scoring Variability

The first and most critical step in stabilizing the bases of readers' judgments is to establish common, explicit scoring criteria. Criteria may either be specified deductively by invoking standards derived from the rhetorical tradition (e.g., Kinneavy, 1971) or inductively by seeking commonality among readers' comments on papers (Diederich, 1974; Freedman, 1978). Systematic training on common scoring criteria has proved to reduce some kinds of interrater variability effectively. (Stalnaker, 1934; Diederich, 1974). As a result of these pioneering studies, standard methodology now includes training of raters on the use of rating scales until a high level of agreement among raters is achieved. In a recent study of the discriminative validity of alternative scoring rubrics, Winters (1978) suggested that high rater reliability coefficients in pilot or in final rating sessions might not necessarily signal standard, uniform interpretation of rating scales over rating occasions and across rater groups. During rater training she observed that less operationalized scale rubrics stimulated extensive discussion and interpretation and suggested that different rater groups might achieve high reliability, but have interpreted vague criteria differently by devising different specific decision rules for the same ambiguous criteria. Thus, high reliability coefficients might be obtained, but at the cost of accurate, replicable scoring. As Winters implies, redefinition of criteria by the social rating group can have serious implications for the fairness of ratings across rater groups.

Rater Drift

Even with training for rater consensus, when raters practice applying explicit criteria, rating fluctuation may still occur. The deviation of raters from previously-shared criteria is termed "rater drift" and may be signaled by lowered inter-rater reliabiliy and differences between raters' criteria interpretation and expert-generated criterion-based ratings.

Rater drift is particularly a problem when there are large sets of papers to be scored. Shifting criteria or drift may be caused by rater fatigue, or by more systematic influences, such as the quality range of the sample of papers being read or idiosyncratically valued criteria. In a description of the rater as a source of error, Braddock et al. (1963), discussed the need for controlling for rater fatigue. They cited fatigue as a cause for raters to become severe or erratic in their evaluation or to place more weight on particularly noticeable essay elements such as mechanics. Godshalk, Swineford, and Coffman (1966) found significant differences between papers scored holistically early and later in a set of 646 papers. Coffman (1971b) warned that even when two sets of scores derive from changing combinations of raters, "there may still be differences in the means and standard deviations attributable to order effects -- that is, the tendency of groups of raters to shift their standards as the reading proceeds" (p. 276). Coffman (1971a) also discussed raters' tendency to regress to their own internalized set of standards and recommended practice on common criteria.

Rater drift impairs the technical quality of rating results by reducing inter- and intra-rater reliability, and more importantly, compromises the validity of ratings. However, writing assessment programs do not seem to acknowledge rater drift as a validity problem, nor do they deal with rater drift directly.

State-of-the-Art Procedures for Treating Scoring Variability

Current rating procedures (Conlan, 1976; Office of the Los Angeles Superintendent of Schools, 1977) generally follow methods recommended by Braddock et al. (1963), and Coffman (1971a) and have evolved a number of methods to deal with rater variability. Typically, raters begin by practicing applying a rubric to a sample set of papers. The nature and relative specificity of scale criteria and scoring formats (*holistic* vs. *analytic*) vary, as do the weights of component criteria. Before independent rating begins, trainers conduct a reliability check. Sometimes consensus is checked statistically; sometimes it is indicated by a show of hands.

During independent ratings, methods for dealing with rater agreement tend to take two tacks: correction and maintenance. Procedures which emphasize correction use *post hoc* methods to treat score discrepancies.

Common options are: 1) having a third reader score any paper where the first readers disagree by more than one point; 2) using the sum of two ratings as a total score; 3) randomizing the order in which two raters score an essay in order to distribute rater error, although often the randomization occurs in a single day. These post hoc correction procedures sidestep the validity problem of the changing criteria employed by the drifting rater.

A second set of procedures for dealing with rating variability aims at maintenance of scoring accuracy. Periodic consensus checks on identical papers are interspersed at varying intervals. Checks may be common to all raters, discussed in the group, discussed within rater pairs or discussed with a "master" rater. In the procedure, discrepancies are called to the rater's attention and their bases revised. These maintenance procedures at least attempt to prevent, detect, and control scoring error by providing feedback to individual raters regarding the accuracy and consistency of their scoring decision rules.

Rating Variability in Competency Assessment Research

In a series of studies examining dimensions important in the formulation of valid, instructionally sensitive writing assessments, we documented the effects of several stringent procedures for attaining and maintaining rater congruence and fidelity to the rating scale. One component of the methodology was to develop analytic scoring rubrics referenced to basic structural features of a discourse mode. Explicit criteria were designed to reference operational, instructionally manipulatable elements of the paper. Raters practiced applying the scoring rubric in intensive training sessions and reliability checks using generalizability statistics were calculated to assure inter-rater reliability. During final, independent ratings, common checks occurred at frequent intervals. Discrepancy resolution procedures were of several types, including group discussion or pair discussion. The research focus of these studies was on variations of the tasks of writing rather than on variables influencing the rating process, yet the accumulating data indicated that stabilizing the judgment process was a complex issue--one deserving direct experimental investigation. This conclusion derived primarily from three of our studies in which we observed rater drift surface as a problem, despite the different procedures used to prevent it. We also began to inspect indices of scale stability by looking at scores given by raters trained at different times to the same set of papers.

Rater Drift

In our writing assessment research our initial scoring concerns were to establish and maintain rater agreement. To determine that this occurred, we compared reliabilities obtained immediately after training (on a pilot test of independent ratings) and after the final ratings. Table 1 presents a comparison of generalizability coefficients marking rater agreement levels on pilot and final ratings.

- - - - - - - - - - -
Insert Table 1 here
- - - - - - - - - - -

The first rating procedure was employed in Study 1 where Spooner-Smith

Table 1

Comparison of Generalizability Coefficients for Rater Agreement Immediately After Training and After Final Ratings

Study 1 - Expository Scale I (Spooner-Smith, 1978)

	F	Dev	O	Su	Pa	M	Total
	GC	GC	GC	GC	GC	GC	GC
Pilot - 4 raters n=15	.94	.92	.94	.83	.94	.80	.90
Final - 2 ratings n = 112	.84	.80	.85	.85	.80	.95	.90

Study 2 - Expository Scale II (Quellmalz and Capell, 1979)

	GI	F	O	S	M	Total
	GC	GC	GC	GC	GC	
Pilot - 4 raters	.74	.63	.74	.77	.73	
Final - 2 ratings	.67	.59	.61	.57	.52	.66

Narrative Scale II

	GI	F	O	S	M	Total
	GC	GC	GC	GC	GC	
Pilot - 4 raters	.86	.76	.79	.76	.52	
Final - 2 ratings	.84	.60	.72	.72	.69	.83

Study 3 - Expository Scale III (Baker and Quellmalz, 1980)

	GI	Gen Comp	Coh	Po	Su	M	Total
	GC	GC	GC	GC	GC	GC	GC
Pilot - 3 raters	.74	.65	.86	.93	.84	.71	.89
Final - 2 ratings	.66	.71	.62	.83	.71	.76	.81

Narrative Scale III

	GI	Gen Comp	Coh	Po	Su	M	Total
	GC	GC	GC	GC	GC	GC	GC
Pilot - 3 raters	.83	.75	.62	.87	.54	.85	.79
Final - 2 ratings	.70	.76	.53	.87	.67	.68	.81

KEY
GC = Generalizability Coefficient

Study 1 (Spooner-Smith, 1978)

F = Focus
Dev = Development
O = Organization
Su = Support
Pa = Paragraphing
M = Mechanics
T = Total

Study 2 (Quellmalz and Capell, 1979)

GI = General Impression
F = Focus
O = Organization
Su = Support
M = Mechanics
T = Total

Study 3 (Baker and Quellmalz, 1980)

GI = General Impression
Gen Comp = General Competency
Coh = Coherence
Po = Paragraph Organization
Su = Support
M = Mechanics
T = Total

(1978) compared direct and indirect measures of writing competence. Four raters received five hours practice applying an analytic rubric, Expository Scale I, to a set of papers representative of the experimental set. The top table presents Spooner-Smith's interrater reliabilities for four raters on the pilot test conducted immediately after training and on the final independent ratings of the experimental papers. During the final independent scoring, raters read, rated and discussed discrepancies on a common paper as a group approximately every hour to check adherence to criteria. While the total score reliability on the final ratings remained high, reliabilities of four of the six subscales dropped as much as .14, indicating some degree of rater drift from original consensus levels.

The second rating procedure occurred in Study 2 (Quellmalz & Capell, 1979) which compared writing performance in different discourse and response modes. Following scale training procedures employed by Spooner-Smith (1978), pilot tests of interrater reliabilities for two revised analytic rubrics, Expository Scale II and Narrative Scale II, checked level of agreement of the four raters prior to final rating. Additional training occurred on any subscale where the generalizability coefficient was less than .70. During final scoring, rater pairs read and discussed common papers after every 20 independent ratings. The two tables for Study 2 indicate, again, that agreement levels on the total scores were acceptably high, but that reliabilities on three of the expository subscales deteriorated as much as —.20. The interpretation of these data was that the frequency and nature of the common check procedures were still not curbing rater drift adequately.

Consequently, Study 3 implemented a revised rating procedure. Study 3 (Baker & Quellmalz, 1980) investigated the effect of modality of topic presentation on eighth grade writing performance. Three raters participated in scale training for analytic Expository Scale III and Narrative Scale III. Following a pilot test of inter-rater reliability, the three raters independently scored the experimental papers. Each paper received two ratings. Common checks occurred every hour and were discussed by the entire group.

As the two tables for Study 3 indicate, agreement levels fall on General Impression, but not on the General Competency rating. Reliabilities plummetted on the expository Coherence ratings and on the Mechanics ratings of the narrative scale. These comparisons of pilot and final reliabilities for Study 3 suggested that the revised checking procedure was generally maintaining rater agreement, but still did not prevent drift on some subscales.

In a more detailed inspection of the emergence of rater drift in Study 3, we also compared reliabilities and mean scores on papers scored early and late in the rating sequence (see Table 2). Table 2 presents the early vs. late comparisons for Expository Scale III and Narrative Scale III. On the expository scale, reliabilities across all rater pairs remain high (α .76 to .85) except on the General Impression and Coherence subscales. Parametric comparisons of mean scores on early vs. late papers did not reach statistical significance, but late scored papers received slightly higher ratings than early scored papers.

Reliabilities on Narrative Scale III remained high on General Compe-

TABLE 2

Comparison of Early vs. Late Scored Papers in Study 3

(Baker and Quellmalz, 1980)

Expository Scale III

	Inter-rater Reliabilities		Mean Scores			
	Early	Late		Early	Late	t
General Impression	α .85	.69	$\bar{X}$	2.28	2.29	.97
			S.D.	1.07	.85	
General Competency	α .75	.77	$\bar{X}$	2.20	2.43	.23
			S.D.	.91	.86	
Coherence	α .78	.57	$\bar{X}$	2.39	2.63	.21
			S.D.	.88	.90	
Paragraph Organization	α .87	.86	$\bar{X}$	2.03	2.22	.40
			S.D.	1.05	1.08	
Support	α .78	.76	$\bar{X}$	2.99	3.11	.51
			S.D.	.85	.90	
Mechanics	α .67	.82	$\bar{X}$	2.18	2.99	-1.08
			S.D.	.85	.76	
Total	α .87	.85	$\bar{X}$	14.78	15.89	-1.06
			S.D.	4.86	4.49	
	n=40	n=40	n=40		n=40	
Narrative Scale III						
General Impression	α .78	.71	$\bar{X}$	2.62	2.19	2.31*
			S.D.	.92	.73	
General Competence	α .81	.78	$\bar{X}$	2.54	2.20	1.84
			S.D.	.87	.78	
Coherence	α .77	.46	$\bar{X}$	2.60	2.31	1.60
			S.D.	.99	.59	
Paragraph Organization	α .93	.85	$\bar{X}$	2.22	2.03	.74
			S.D.	1.29	1.00	
Support	α .84	.84	$\bar{X}$	2.82	2.51	1.68
			S.D.	.97	.68	
Mechanics	α .68	.80	$\bar{X}$	2.30	2.16	.82
			S.D.	.80	.74	
Total	α .90	.86	$\bar{X}$	14.35	13.03	1.49
			S.D.	4.94	3.44	
	n=40	n=50	n=40		n=50	

α = alpha coefficient

*$p < .05$

tence, Support, Mechanics and Total score. General Impression reliability dropped .08, Coherence dropped substantially (α .77 to .46) and Paragraph Organization fell (α .93 to .85). Contrasts of mean differences between early and late scored narrative papers revealed a significant difference on General Impression ratings. Papers scored later received lower ratings than those scored earlier. All subscale scores were lower for late scored papers. These findings are consistent with other research (Godshalk et al., 1966) that reported raters became more severe as scoring progressed. In Study 3, Expository papers were scored before Narrative papers, so late scored Narrative papers were at the very end of the entire scoring sequence.

Inspection of the scoring data from the three studies suggests that rater drift within a scoring session can occur and weaken scoring rigor. Raters' judgments waivered on some subscales more than others, signalling a need for more careful explication of criteria on those subscales and practice on their application. Since state-of-the-art procedures for controlling rater drift were employed and even refined in these studies, the data implied the need to continue to examine methodologies for detecting and preventing rater drift.

Scale Stability

A validity concern coordinate with maintenance of scale fidelity within rating occasion is assurance of judgment accuracy across rating occasions. Standards of fairness and methodological rigor mandate that criteria apply uniformly across sets of raters and sets of papers.

Prevailing practice does not seem to recognize stability as a technical problem. Large scale assessments do not routinely report and inspect a series of rater reliabilities for separate scoring sessions. Even reliability indices are not sufficient, however. Comparisons of mean scores on common papers should supplement reliability statistics. Scale stability could be demonstrated by comparing scores on a common set of papers given by different rater sets trained separately, or by comparing scores from the same raters rating at different occasions. While we have not yet investigated this phenomenon within an experimental paradigm, we have, however, inspected scoring data gathered during the process of our other writing assessment research in an attempt to understand the nature of variables influencing scale stability.

Our Table 3 presents the means and standard deviations of essay scores given by two different rater sets to the same papers. Raters A and B scored 30 expository essays. Rater pairs 1, 2 and 3 rated these same 30 essays in the course of Study 3. Rater pairs 1, 2 and 3 were using Expository Scale III, a revision of the analytic expository rating scale used by Raters A and B. Therefore only scores from those subscales that were not significantly changed were entered into the analysis. Agreement levels were not calculated due to the small sample size.

Inspection of the means reveals that Raters A and B gave generally higher ratings than Rater pairs 1, 2 and 3. Comparisons of means for each subscale and the total score were all significant. While the small number of papers clearly limit interpretation of these data, they do document that criteria definition and application did change from one rating session to the next.

Table 3

Comparison of Essay Scores* Given by Different Rater Sets on Separate Occasions

Subscales		Ratings			
		Occasion 1 Raters A and B	Occasion 2 Raters 1-6	t	df
General Competence	$\bar{X}$	2.92	1.65	2.77*	57
	s.d.	.62	.38		
	n	29	30		
Paragraph Organization	$\bar{X}$	2.19	1.46	3.67*	58
	s.d.	.98	.50		
	n	29	31		
Support	$\bar{X}$	2.76	2.07	3.25*	59
	s.d.	1.08	.50		
	n	29	32		
Total	$\bar{X}$	11.81	8.97	4.38*	59
	s.d.	3.17	1.76		
	n	29	32		

---

*Scores by rater pairs 1-6 were transformed from a score range of 1-6 to 1-4 to permit analyses.

In addition to looking at the scores different raters trained at separate occasions gave to the same set of papers we inspected intra-rater agreement of scores a pair of raters gave to common papers scored at different sessions. Table 4 displays means and standard deviations of a rater pair (N) which participated in two different rating sessions.

- - - - - - - - - -

Insert Table 4 here

- - - - - - - - - -

In Study I, rater pairs M and N scored essays from a general high school population which were then "salted in" a set of college admission essays read for Study 2. In Study 2, pair N read the eight essays they had scored previously in Study 1 and 8 additional essays from that study that they had not personally scored. The means of pair N in the two studies are fairly comparable except on Support and Mechanics. In contrast, the means of pairs M and O are substantially different. Pair O means are consistently lower. The greater stability of means for pair N may suggest that they were applying criteria in a uniform manner. Pair O was probably influenced by the overall higher quality of the college admissions sample, thus making the "salted in" general population high school seem worse. Methods for eliminating this subtle "norming" of presumably explicit criteria to the quality range of particular sample is a phenomena requiring further research.

Our intent in inspecting these admittedly limited data was to illustrate one method for tracking the stability of rating scale application. Writing assessments could systematically include a "check" set of papers in each rating session to document the comparability of judges' decision

TABLE 4

Comparison of Rater Pair Scores Across Studies

		Study 1		Study 2	
Rater Pair		M	N	N	O
CSE Subscale					
General Impression	$\overline{X}$	1.92	1.28	1.00	.94
	s.d.	1.32	1.37	1.13	.91
	n	6	8	16	16
Focus	$\overline{X}$	2.08	1.71	1.69	1.53
	s.d	.38	.9	.48	.50
	n	6	8	16	16
Organization	$\overline{X}$	2.33	1.65	1.72	1.38
	s.d.	.98	.6	86	.50
	n	6	8	16	16
Support	$\overline{X}$	2.42	2.76	2.00	1.63
	s.d.	.92	1.15	1.78	.50
	n	6	8	16	16
Mechanics	$\overline{X}$	2.50	2.20	1.78	1.75
	s.d.	.84	.70	.77	.58
	n	6	8	16	16
Total	$\overline{X}$	11.25	9.60	8.19	7.21
	s.d.	3.71	2.79	3.40	2.53
	n	6	8	16	16

rules at different rating sessions. We believe that scale stability across topics, quality range of papers and sets of raters can be achieved and that the factors influencing scale stability require systematic investigation.

Summary and Recommendations

The need for stabilizing the scoring process is critical to the validity of writing assessments. Direct evidence of student writing competence, actual written production, is a necessary condition for content and construct validity; it is not sufficient, however. Rater's judgments must be replicable and defensible. We believe that explicit rating criteria are a condition for defensibility and replicability. Our rater drift comparisons suggest that total scores and a holistic score seem to mask fluctuations in judgments on the elements that contribute to the more global summary scores. We suspect that, at least during scale development and validation, assessments should collect separate ratings on component text features such as Support and Coherence that contribute to a total score. Otherwise, there is no way to identify and track consistency of the bases for global judgments.

Certainly, scale training and an initial reliability check is essential. Rather than relying primarily on randomization or statistical procedures to correct for rater drift *post hoc*, rating methods should intersperse periodic checks into lengthy, independent scoring. The variables making these checks effective for maintaining agreement and scale fidelity require further investigation. Frequency of checks is one important factor; the nature of feedback on scoring accuracy is even more essential. We are currently conducting research on methods for curbing rater drift.

Scale stability is a critical validity issue for competency-based writing assessment. Large scale assessments can, at least, document stability by tracking scoring of a core set of papers by different groups of raters. Methodologies for selecting and preventing scale instability should also receive direct experimental attention. Fair, informative, generalizable, defensible scoring procedures are necessary requirements of sound writing assessment.

References

Baker, E. L., & Quellmalz, E. S. Issues in eliciting writing performance: Problems in alternative prompting strategies. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, April 1980.

Braddock, R., Lloyd-Jones, R., & Schoer, J. Research in written composition. Urbana, Ill.: National Council of Teachers of English, 1963.

Coffman, W. E. Essay Examinations. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D. C.: American Council of Education, 1971a.

Coffman, W. E. On the reliability of ratings of essay examinations in English. Research in the Teaching of English, Vol. 5(1), Spring 1971b.

Conlan, G. How the essay in the CEEB English test is scored. Princeton, N. J.: Educational Testing Service, 1976.

Crowhurst, M. Syntactic complexity in narration and argument at three grade levels. Canadian Journal of Education, 1980.

Diederich, P. B., French, J. W., & Carlton, S. Factors in judgments of writing ability. Princeton, New Jersey: Educational Testing Service, 1961.

Diederich, P. B. Measuring growth in English. Urbana, Ill.: National Council of Teachers of English, 1974.

Eels, W. C. Reliability of repeated grading of essay-type examinations. Journal of Educational Psychology, 1930, 21.

Findlayson, D. S. The reliability of the marking of essays. British Journal of Educational Psychology, 1951, 21, 126-134.

Follman, J. C., & Anderson, J. A. An investigation of the reliability of five procedures for grading English themes. Research in the Teaching of English, 1967, 190-200.

Freedman, S. How characteristics of student essays influence teachers' evaluation. Journal of Educational Psychology, 1978, 70.

Godshalk, F. E., Swineford, F., & Coffman, W. E. The measurement of writing ability. New York: College Entrance Examination Board, 1966.

Kinneavy, J. R. A theory of discourse. In the Aims of Discourse. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1971.

McColly, W. What does educational research say about the judging of writing ability? The Journal of Educational Research, 64, No. 4, December 1970.

Office of the Los Angeles County Superintendent of Schools. A common ground for assessing competencies in written expression, review copy. Los Angeles: Division of Curriculum and Instructional Services, 1977.

Pitts, M. The relationship of classroom instructional characteristics and writing in the descriptive/narrative mode. Report to the National Institute of Education, Los Angeles: UCLA Center for the Study of Evaluation, 1978. (Grant No. OB-NIE-G-78-0213)

Prater, D., & Padia, W. Effects of modes of discourse in writing performance in grades four and six. Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

Quellmalz, E. S. Interim report. Defining writing domains: Effects of discourse and response mode. Center for the Study of Evaluation, University of California, Los Angeles, 1979.

Quellmalz, E. S., & Capell, F. Defining writing domains: Effects of discourse and response mode. Report to the National Institute of Education, November, 1979. (Grant No. OB-NIE-G-78-0213 to the Center for the Study of Evaluation)

Spooner-Smith, L. Investigation of writing assessment strategies. Report to the National Institute of Education, November 1978. (Grant No. OB-NIE-G-78-0213 to the Center for the Study of Evaluation.)

Stalnaker, J. The construction and results of a twelve-hour test in English composition. School and Society, 1934, 39.

Winters, L. The effects of differing response criteria on the assessment of writing competence. Grant No. OB-NIE-G-78-0213, Los Angeles, California: Center for the Study of Evaluation, November 1978.

1.

EFFECTS OF VISUAL OR WRITTEN TOPIC INFORMATION ON ESSAY QUALITY

Eva L. Baker and Edys Quellmalz

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

When one considers the critical properties a test must have, the element of validity is almost always discussed. Validity means the test measures what it is supposed to, and, as an essential corollary, makes that measurement fairly. Fairness or equity involves the kind of chances students have to demonstrate their competency. In some cases, test equity is also inferred from the shape of the resulting distribution of scores for different student groups. The task of writing assessment is a particular and special subset of the testing process. Writing assessment contains three special features that increase the importance of its research: 1) Writing skill is recognized by the public and by educators as a critical goal of schooling; 2) the study of writing provides access to the study of cognitive processes; 3) writing assessment serves as a case from the larger set of constructed responses in achievement testing, a set to which relatively little psychometric theory has been applied.

Although the use of actual student writing samples seems the most obvious way to obtain estimates of student performances, it is loaded with practical complexities. Unlike multiple choice tests, which appear to be fixed artifacts, essay-based assessment appears less constrained in the writing tasks themselves, particularly in the actual directions or prompts given to learners and the criteria used to score performance. Of course, differences in either of these dimensions greatly affect the inferences we make. Were writing assessment to remain the domain of individual teachers, as they privately teach and assess, we would expect that idiosyncracies of tasks and scoring schemes in particular classrooms would be balanced over time by the number of different teachers to whom any given student was exposed. Yet, writing assessment assumes public rather than private proportions, as exemplified by competency testing for high school graduation and statewide assessment programs. The public functions of writing assessment bring with them the necessity to develop and to display to students and teachers the specifications guiding the preparation of writing tasks.

From a research perspective, moreover, writing assessment presents a special opportunity to understand the learning process. By studying information requirements, the effects of cues and supports given students, the technical aspects of assessment can be improved and desirable features identified for inclusion in writing instruction. Writing, as much as any school-trained activity, shows us how students think, how they organize information, and how they understand subject matter.

Our work on writing assessment related to a general assessment framework. This framework includes elements relating to 1) social and intellectual motivation; 2) student ability and information; 3) features of tasks; 4) criteria used for scoring. (See Figure 1) Although these categories could be elaborated almost infinitely, our research focuses on elements endogenous to the writing tasks, such as cues, modes of discourse, and information.

- - - - - - - - - - - -

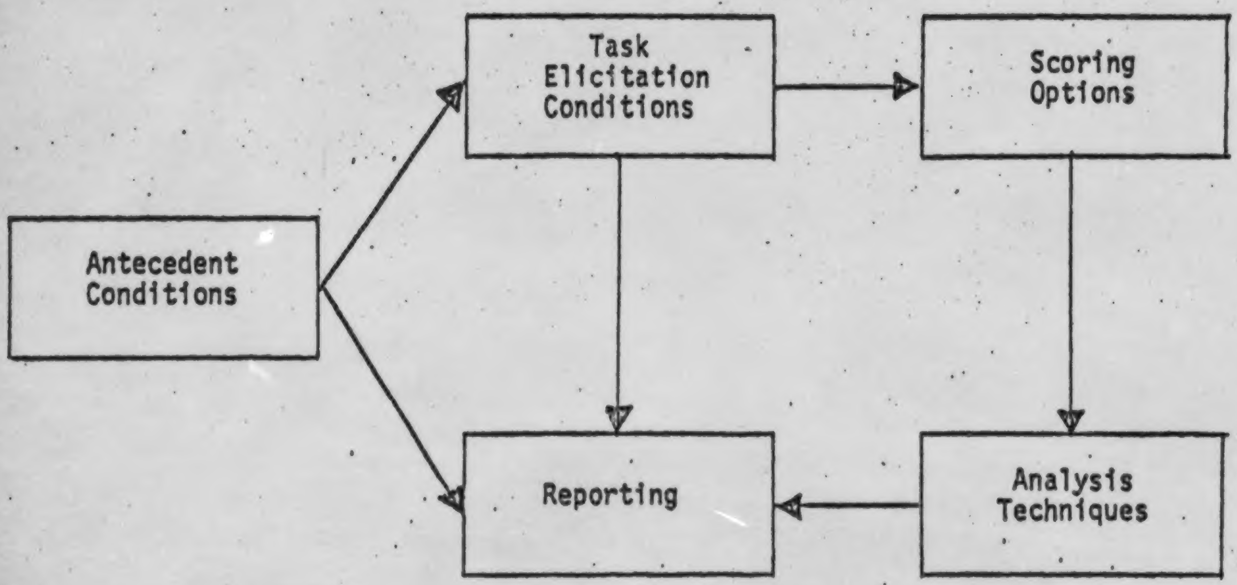
Insert Figure 1 here

- - - - - - - - - - - -

Tasks for Writing Assessment

The first problem of writing assessment is the selection of a topic,

CONCEPTUAL FRAMEWORK: TEST DESIGN


Task Elicitation Conditions
Scoring Options
Antecedent Conditions
Reporting
Analysis Techniques

Information System Linking Testing and Instruction

Figure 1

direction, or prompt to elicit student composition. It is common to give assignments on topics that students have some information about but which avoid systematically advantaging particular students with specialized knowledge. This "nodding acquaintance" mode of topic identification results in bland, general topics, such as "My Street" -- topics unlikely to generate real enthusiasm and meaning for students. Nonetheless, particular students are thought to be about equally prepared on such topics. In addition, these inoffensive assignments will not disturb parents or school administrators. The "general" experience tapped by such tasks corresponds to the general "frames" described by Minsky (1975) and Anderson (1972) in research on the cognitive aspects of reading comprehension. These frames consist essentially of certain general information and common referents prerequisite to the ability to use special tactics to generate responses. In particular, the reliance on common frames or referent "general experience" becomes more and more risky as students come from more diverse backgrounds. With less shared experiences among students, writing tasks themselves may need to provide information for students to write about. We are led to the simple question of whether possession of specific information affects students' ability to write. In other words, can students demonstrate their writing ability when they have little specific information to convey? Or, is writing fluency independent of specific information and related more to the general or common referent information? Most importantly, how can students be supported so that they can display their writing competency at its highest level?

To explore these questions, we need to provide specific information to students so that they may have "content" for composing. How can one go about this task? English composition teachers have attempted to provide sufficient information to learners to enable them to write knowledgeably and productively. Such efforts most often take the form of an extended set of written directions which sets the context and audience and provides a brief discussion of the purpose of the essay. One limitation of this form of prompt, the "extended written passage," is the reading comprehension load placed upon the would-be writer, a particular disadvantage for a large number of poor readers. Another consequence frequently noted by teachers is imitation. Students may mimic the form and style of the extended prompt itself.* The strongly styled prose instructions may jeopardize students' writing and, consequently, the accuracy of the assessment effort.

Because pictures convey information that is both general and specific, they have been used formally in some assessment settings as a way around excessive reading burden. In fact, Levin and Lesgold (1978) describe the facilitation properties of pictures in reading comprehension tasks. If pictures enhance comprehension differentially for poor comprehenders, then the use of picture prompts as substitutions or elaborations for prose directions should positively affect writing performance, particularly for students with poorer writing skills.

Overview

The study represents a start in unraveling the relationships between information in writing tasks and writing performance. In this research, the information provided in picture prompts for eighth-grade writing tasks which were varied in two modes of discourse, exposition and narrative, was compared with written prompts. Although pictures tend to be associated with narrative or descriptive writing, the frequency of and generally dismal performance on expository tasks led us to test picture prompts in expository tasks as well as descriptive writing. An experiment was conducted where eighth grade students were randomly assigned to receive either a picture or written prompt and either an expository or narrative writing task. Students also completed a test of reading ability. Results were assessed using both general and analytic scoring schemes and were examined for students of different reading achievement levels.

---

*The phenomenon is much like a tendency to write short, zippy declarative sentences after reading Hemmingway, convoluted and sentimental passages after reading Dickens, or to forego the rules of capitalization after confronting e. e. cummings.

Subjects and Sampling

Students were sampled as part of an evaluation of eighth grade achievement in a study of California educational reform. The eighth grade level was chosen because, at that grade, students often write both narrative and expository compositions. Eighteen schools were sampled, to represent a scheme stratified by school size, percentage of low or non-English speaking students, California geography, and socioeconomic status (SES). Two heterogeneously grouped classrooms in each school at the eighth grade were randomly selected.

Instrumentation

Writing Task. Because the different information conveyed by picture and prose is a matter of aphorism, the study made no attempt to equate concrete or general bits of information in these two stimulus classes.

Within each classroom, students received one of four treatments: 1) picture prompt/directions for exposition (PE); 2) picture prompt/directions for narration (PN); 3) written prompt/for exposition (WE); 4) written prompts/for narration (WN).

Reading Test. The reading test consisted of 68 items and was composed of three subscales: vocabulary (21 items), literal comprehension (24 items), and inferential comprehension (23 items). The tests were used to assess reading skills in a statewide evaluation study. Tests were generated using domain-referenced testing procedures (see Hively, 1974; Herman, 1977; Baker, 1977). Spearman-Brown coefficients were computed and coefficients .76, .80 and .72 were obtained respectively, for the three subscales. The Spearman-Brown coefficient for the total 68 item test was .92. The difficulty by subscale was .64 for vocabulary, .65 for literal comprehension, and .64 for the inferential comprehension subscale..

Scoring Systems for Dependent Measures. In previous studies, (Winters, 1978; Spooner-Smith, 1978), the utility of alternative scoring strategies, e.g., holistic and analytic procedures, was assessed using high school and college populations. Since this sample population was younger, a pilot study of scoring procedures was conducted using thirty papers, selected at random from the entire sample. On the basis of this pilot study (Baker & Quellmalz, 1979), the use of T-units as a dependent measure was excluded for this research. A general impression (GI) or holistic assessment was given to each essay score, using a six-point scale in the pilot study. In addition, an analytic scoring rubric was applied to each essay on schema previously employed in other writing research (Pitts, 1978; Quellmalz, 1979; Winters, Spooner-Smith, op. cit.). This rubric consists of four subscales (of six-point range) on structural features of the essay: 1) paragraph organization, 2) coherence, 3) support and 4) mechanics. These subscales are combined into a General Competency scale. Previous work (Quellmalz & Capell, 1979) identified the relative independence of the scales.

Socio-economic Indicators. The State of California has no direct index of SES for its secondary school populations. Instead, the percent of students receiving Aid to Families with Dependent Children (AFDC) is used as a proxy.

Procedures

Teachers were directed to administer the test of reading comprehension on the first day of testing during the spring. On the second day, students received a writing task to complete in forty minutes. The four treatments were randomly assigned within classroom. Teachers returned by pre-paid mail coded student response booklets to the researchers.

Rater Training. Three pairs of raters were trained to use this rubric on expository and narrative sample papers.* Six hours were spent training and practice scoring in the expository mode until an acceptable level of agreement was reached (alpha = .83, generalizability coefficient = .89). All expository papers were then rated in a group. During a subsequent session, practice was provided in applying the rubric to narrative papers until acceptable concordance was reached (alpha = .83, generalizability coefficient .79). Raters first gave each paper a general impression rating, which included estimates of the style, creativity, structural and mechanical features of the paper. Next, they gave a general competency score with regard to the subscales - coherence, support, paragraphing, and mechanics. Last, the raters scored each subscale. All scores were assigned from a "1" to "6" scale.

---

*Student paper length varied from one-half page to two pages. Student papers were not retyped for scoring.

Analysis

Data were returned from thirteen of the eighteen schools. Because of constraints about information for individual students in this sample, only school level data were available regarding SES and language dominance. According to a comparison at the sample means with the statewide means, the attrition (the loss of five schools) occurred in those schools with lower SES and higher percentages of low English speakers. The distribution of AFDC in our total sample was 10% and in the returned sample 6%, indicating that the drop-out took place in low SES schools. In addition, the percentage of low English speaking students was reported as less than one percent for our sample, much lower than for the entire California population. Clearly, the lower achieving side of the sample did not return the measures. Although the school level/student level SES data problem was clear, and we had hoped that some indication of SES might assist us in our analysis, we found that our test of reading achievement correlated .62 with AFDC. This correlation corresponds to relationships found over the last few years of studies of achievement and SES in California elementary schools (Baker, 1976; Baker & Herman, 1977; Baker, Herman & Yeh, 1978; Quellmalz & Baker, 1981).

Experimental Contrasts. Students' writing performance was assessed using 2x2x2 design, employing two levels of reading scores, high and low split at the mean, two types of prompt, and two types of discourse task.

Data were separately analyzed for the General Impression scale (GI), for the General Competency scale, for the four subscales of the analytic rubric, and for the total. Means and standard deviations for each variable were computed and multiple classification analyses of variance were performed.

Results

Means and standard deviations of the dependent measures by blocking factors are presented in Tables 1 through 24. Because of the large number of analyses conducted, some significant findings would be detected by chance alone. Only those findings which show up consistently across measures will be discussed.

Overall Findings. The salient feature of the tables is the relatively poor performance of students in writing competence. Raters were instructed to use "4" as a score of sufficient competency or, oxymoronically, minimum-mastery. No average, either for high scoring readers or for any treatment variation, is 4 or higher. This finding is particularly depressing in the light of the drop-out analysis and the inference that schools with poorer performing students did not return the measures for scoring.

Holistic Scoring. Two forms of holistic scoring were used: the General Impression (GI) scale, which included style and other "intangibles" such as creativity, and the General Competency score, an estimate of the total "competency" of the paper with regard to the features of the analytic scale. Recall that these two scores were given before detailed analytic scoring took place. Using the total reading score as a blocking factor,

- - - - - - - - - - - - -

Insert Tables 1-8 here

- - - - - - - - - - - - -

a significant two-way interaction was found on GI for mode of discourse and reading ability (F = 4.17, df = 1,212, p = .04). Inspection of Table 1 suggests that performance by the low readers on narrative tasks was particulary poor. With regard to the General Competency score, mode of discourse and prompt form significantly interacted (F = 4.50, df = 1,212, p = .04) with the inferential comprehension blocking factor (see Table 1) and missed by a little (F = 2.58, df = 3,212, p = .055) on all other blocking runs. A speculative interpretation is that written expository and picture narrative combinations are most facilitative on General Competency. Inspection of Tables 1-9 suggests that these effects were true for the better readers, a speculation supported by the three-way interaction in Table 7, (F = 4.08, df = 1,212, p = .045). The findings for the Coherence dependent measure, (Tables 9-12) display a significant main effect for prompt type (F = 6.12, df = 1,212, p = .01) in favor of pictures. This finding is also replicated on the vocabulary subscales (F = 5.92, df = 1,212, p = .016). Table 11 provides an interesting case where the poor readers out perform the good readers under the picture-expository condition by about .5 standard deviation, out perform low reading groups without pictures by a greater margin, and about equal the high reading groups who do not have

- - - - - - - - - - - - -

Insert Tables 9-12 here

- - - - - - - - - - - - -

General Impression

Table 1

Total Reading

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High	X	2.36	2.92	2.51	2.53
Total	sd.	0.80	0.89	1.07	0.61
Reading	n=	33	30	35	29
Low	X	2.18	2.06	2.12	1.94
Total	sd.	0.61	0.47	1.08	0.63
Reading	n=	19	24	20	23

Table 2

Inferential Comprehension

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High	X	2.34	2.83	2.55	2.47
Total	sd.	0.81	0.89	1.05	0.61
Inferential Comprehension	n=	28	30	30	33
Low	X	2.25	2.17	2.20	1.92
Total	sd.	0.66	0.62	1.10	0.67
Inferential Comprehension	n=	24	24	25	19

Table 3

Comprehension

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High	X	2.34	2.85	2.59	2.46
Total	sd.	0.79	0.89	1.11	0.64
Comprehension	n=	28	26	29	25
Low	X	2.25	2.25	2.17	2.09
Total	sd.	0.68	0.70	1.02	0.68
Comprehension	n=	24	28	26	27

Table 4

Vocabulary

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High	X	2.44	2.870	2.79	2.42
Total	sd.	0.79	0.967	1.17	0.58
Vocabulary	n=	23	27	24	26
Low	X	2.19	2.20	2.08	2.12
Total	sd.	0.69	0.54	0.91	0.75
Vocabulary	n=	29	27	31	26

General Competency

Table 5

Total Reading

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High Total Reading	X	2.39	2.88	2.59	2.47
	sd.	0.66	0.88	1.03	0.61
	n=	33	30	35	29
Low Total Reading	X	2.13	1.96	2.23	1.94
	sd.	0.62	0.49	0.91	0.76
	n=	19	24	20	23

Table 6

Inferential Comprehension

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High Total Inferential Comprehension	X	2.32	2.75	2.63	2.44
	sd	0.63	0.91	1.04	0.62
	n=	28	30	30	33
Low Total Inferential Comprehension	X	2.27	2.13	2.24	1.87
	sd	0.69	0.66	0.90	0.76
	n=	24	24	25	19

Table 7

Comprehension

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High Total Comprehension	X	2.67	2.79	2.72	2.42
	sd	0.63	0.85	1.08	0.69
	n=	28	26	29	25
Low Total Comprehension	X	2.33	2.18	2.15	2.06
	sd	0.69	0.77	0.80	0.73
	n=	24	28	26	27

Table 8

Vocabulary

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High Total Vocabulary	X	2.46	2.815	2.79	2.44
	sd	0.67	0.97	1.05	0.61
	n=	23	27	24	26
Low Total Vocabulary	X	2.17	2.13	2.19	2.02
	sd.		0.57	0.87	0.78
	n=	29	27	31	26

Coherence

Table 9

Total Reading

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High Total Reading	X	2.59	3.02	2.59	2.43
	sd	0.70	0.98	0.97	0.55
	n=	33	30	35	29
Low Total Reading	X	2.66	2.33	2.35	2.13
	sd	0.55	0.58	1.09	0.51
	n=	19	24	20	23

Table 10

Inferential Comprehension

		Picture		Narrative	
		Expository	Narrative	Expository	Narrative
High Total Inferential comprehension	X	2.52	2.95	2.62	2.46
	sd.	0.73	0.99	1.02	0.51
	n=	28	30	30	33
Low Total Inferential Comprehension	X	2.73	2.42	2.36	2.03
	sd.	0.53	0.64	1.00	0.51
	n=	24	24	25	19

Table 11

Comprehension

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High Total Comprehension	X	2.48	3.04	2.72	2.40
	sd	0.71	0.87	0.96	0.48
	n=	28	26	29	25
Low Total Comprehension	X	2.77	2.41	2.25	2.20
	sd	0.54	0.80	1.02	0.59
	n=	24	28	26	27

Table 12

Vocabulary

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High Total Vocabulary	X	2.70	2.96	2.69	2.31
	sd	0.75	0.94	1.07	0.45
	n=	23	27	24	26
Low Total Vocabulary	X	2.55	2.46	2.36	2.29
	sd	0.56	0.77	0.95	0.64
	n=	29	27	31	26

pictures. Obviously the data are only exploratory, but such findings, if replicated, would suggest a compensatory role for picture-simulated expository writing. The three-way interaction, significant beyond the .01 level (F = 7.87, df = 1,212), suggests a disordinal relationship where pictures facilitate the high readers' narrative production and affect the low readers' expository performance. A mode of discourse by inferential comprehension interaction is also significant (F = 4.50, df = 1,212, p = .035) and suggests that poor readers have less success with narrative tasks. (See Table 10.) No differences were detected on the paragraphing subscale.

For the Support subscale, where use of examples and details are assessed, the findings are the most consistent. With the total reading score as a blocking factor (Table 13), main effects are found for prompt form (F = 21.32, df = 1,212, p = .0001), for mode of discourse (F = 11.96, df = 1,212, p = .001), and a mode x prompt interaction (F = 5.02, df = 1,212, p = .026). In Table 13, we are struck with findings that suggest, with pictures, low readers perform equivalently to high readers in the expository mode and superior to high readers in any other treatment. In addition, it appears that readers make special use of pictures in the narrative mode. This pattern of findings is repeated with Inferential Comprehension scores as a blocking factor and an additional two-way interaction between mode of discourse and reading level is found. Again, these findings support the "special" use good readers make of pictures in the narrative mode (Table 14).

- - - - - - - - - - - - - -

Insert Tables 13-16 here

- - - - - - - - - - - - - -

Support

Table 13

Total Reading

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High	X	3.46	3.13	2.86	2.71
Total	sd	0.65	0.91	0.88	0.61
Reading	n=	33	30	35	29
Low	X	3.40	2.42	2.50	2.41
Total	sd	0.66	0.57	0.86	0.72
Reading	n=	19	24	20	23

Table 14

Inferential Comprehension

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High	X	3.47	3.12	2.83	2.73
Total	sd	0.66	0.91	0.87	0.61
Inferential Comprehension	n=	28	30	30	33
Low	X	3.42	2.44	2.60	2.32
Total	sd	0.65	0.60	0.89	0.69
Inferential Comprehension	n=	24	24	25	19

Table 15

Comprehension

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High	X	3.47	3.12	2,83	2.76
Total	sd	0.58	0.85	0.90	0.61
Comprehension	n=	28	26	29	25
Low	X	3.42	2.54	2.62	2.41
Total	sd.	0.73	0.76	0.86	0.68
Comprehension	n=	24	28	26	27

Table 16

Vocabulary

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High	X	3.52	3.11	3.02	2.65
Total	sd.	0.61	0.92	0.93	0.58
Vocabulary	n=	23	27	24	26
Low	X	3.36	2.52	2.50	2.50
Total	sd.	0.68	0.66	0.79	0.75
Vocabulary	n=	29	27	31	26

The Mechanics subscale provides a different sense of the treatment

- - - - - - - - - - - - - -
Insert Tables 17-20 here
- - - - - - - - - - - - - -

effects. Under all blocking conditions, a main effect is found for mode of discourse. Tables 17 to 20 consistently display an association of the narrative mode with poorer use of mechanics (spelling, syntax, punctuation). With the Inferential Comprehension blocking factor (Table 18), prompt form is significant, favoring written prompts. Perhaps the presence of pictures encourages rapid, sloppy execution of sentence structure.

- - - - - - - - - - - - - -
Insert Tables 21-24 here
- - - - - - - - - - - - - -

On the total writing score, composed of both holistic and analytic scoring procedures, a reading level by mode of discourse interaction effect is found (F = 4.38, df = 1,212, p = .038) and that finding is evidenced either significantly or marginally (p = .06) in the other blocking analyses.

Summary

With caveat underscored, the summary of these data are as follows:

1. Sampled eighth grade children's writing ability, whether scored holistically or analytically, stimulated by picture or written prompt, and with either a narrative or expository task, is below minimal levels of competency.
2. Picture prompts generally facilitate writing, particularly for those subscales which emphasize content detail and coherence.
3. Picture prompts differentially facilitate good and poor readers'

Mechanics

Table 17

Total Reading

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High	X	3.08	2.63	3.15	2.41
Total	sd	0.79	0.91	0.92	0.54
Reading	n=	33	30	35	29
Low	X	2.78	1.92	2.78	1.89
Total	sd	0.70	0.58	0.94	0.71
Reading	n=	19	24	20	23

Table 18

Inferential Comprehension

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High	X	3.03	2.52	3.21	2.36
Total	sd.	0.80	0.90	0.98	0.56
Inferential Comprehension	n=	28	30	30	33
Low	X	2.90	2.06	2.79	1.87
Total	sd	0.73	0.74	0.84	0.72
Inferential Comprehension	n=	24	24	25	19

Table 19

Comprehension

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High	X	3.00	2.54	3.24	2.36
Total	sd	0.84	0.86	1.00	0.64
Comprehension	n=	28	26	29	25
Low	X	2.93	2.11	2.77	2.02
Total	sd	0.69	0.81	0.80	0.66
Comprehension	n=	24	28	26	27

Table 20

Vocabulary

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High	X	3.14	2.54	3.36	2.42
Total	sd.	0.80	0.98	0.86	0.56
Vocabulary	n=	23	27	24	26
Low	X	2.83	2.09	2.75	1.94
Total	sd.	0.72	0.65	0.92	0.68
Vocabulary	n=	29	27	31	26

Total

Table 21

Total Reading

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High Total Reading	X	16.27	17.10	15.76	14.97
	sd	3.62	4.99	5.16	3.08
	n=	33	30	35	29
Low Total Reading	X	15.10	12.35	14.08	12.11
	sd	2.51	2.36	5.03	3.73
	n=	19	24	20	26

Table 22

Inferential Comprehension

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High Total Inferential Comprehension	X	16.06	16.57	15.89	14.73
	sd.	3.60	5.12	5.14	3.14
	n=	28	30	30	33
Low Total Inferential comprehension	X	15.58	13.02	14.27	11.92
	sd.	2.92	3.12	5.08	3.84
	n=	24	24	25	19

Table 23

Comprehension

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High Total Compre-hension	X	15.80	16.86	16.25	14.88
	sd	3.63	5.08	5.37	3.56
	n=	28	26	29	25
Low Total Compre-hension	X	15.89	13.23	13.92	12.61
		2.89	3.47	4.69	3.61
	n=	24	28	26	27

Table 24

Vocabulary

		Picture		Written	
		Expository	Narrative	Expository	Narrative
High Total Vocabulary	X	16.68	16.59	16.98	14.65
	sd	3.69	5.31	5.67	2.97
	n=	23	27	24	26
Low Total	X	15.18	13.39	13.74	12.75
	sd.	2.81	3.33	4.24	4.04
		29	27	31	26

performance conditioned by mode of discourse. Pictures improve poor readers' expository writing to the extent that they at least equal and sometimes outperform good readers in the same treatment condition, exceed good readers in any other condition, and surpass other poor readers. The size of these effects ranges from around 1.5 s.d. to .5 s.d.

4. Picture prompts facilitate good readers' performance on narrative writing tasks.
5. The effect of pictures is negative only for the subscale dealing with mechanics, e.g., spelling, punctuation, etc.
6. Narrative modes receive poorer scores in general, but are particularly hard for poor readers.
7. Expository writing, stimulated by written prompts, provides an adequate opportunity for good readers to demonstrate their competence.

Implications For These Findings

Of most interest, of course, is the replication of these findings under conditions which sample a range of narrative and expository tasks in picture and written prompted situations. The collection of individual demographic data would also allow for finer grained interpretation.

The analysis of why the narrative mode fares less well than expository may be attributed to long term practice effects. Children may write more expository than narrative prose, despite contrary claims of curriculum guides. It is surely the case that raters have more practice, and comfort, with expository rather than narrative writing. Thus, the practice effects of raters may be perpetually confounded with those of students.

What is so powerful about picture prompted expository writing for poorer readers? Pictures appear to provide the necessary content for students to write about. Perhaps poor writing performance results from the lack of a content repertoire to write about. Students may be induced to express ideas if they are presented with a content base. Similarly, one might question why good readers seem to do well with narrative when stimulated by pictures. In our studies, the narratives involved "making up" a story and called for some generative behavior of both content and form. At one level, this task is much more abstract than that of expository writing, since a narrative line needs movement, imagination, and specific content. It may depend more on the general "frame," a knowledge of standard story structures, and perhaps the experience of hearing narratives read aloud. Poor readers may not have the skill to "make up a story" and a single picture may present an insufficient prompt for them. Its effect may actually be distracting.

It is educators' penchants to sound the alarm for individualized instruction whenever disordinal interactions occur. But our task is assessment, rather than exclusively instruction. Instead of matching good or poor readers to one or another combination of prompt forms and modes of discourse, our responsibility may be to provide students with alternative opportunities to demonstrate their writing competency. Apart from speculation of the function of prompts, our data suggest that single, arbitrary writing prompts do great disservice, particularly to those students who may need our special attention.

Perhaps the greatest challenge is to find ways to help students to retrieve and use the content that they already have to write about but do not recognize or acknowledge. They may be able to write, if we give them something to say. The instructional implications of this analysis would suggest we spend a great deal more time and care in "pre-writing" activities to assure students have something to communicate.

References

Anderson, R. C. How to construct achievement tests to assess comprehension. Review of Educational Research, 42(2), 1972, p. 140-170.

Baker, E. L. Long range plan, 1978-1982. Center for The Study of Evaluation, University of California, Los Angeles, 1977.

Baker, E. L. The evaluation of the early childhood education program. Los Angeles, CA: Center for the Study of Evaluation, 1976.

Baker, E. L., & Herman, J. Early childhood education. Los Angeles, CA: Center for the Study of Evaluation, 1977.

Baker, E. L., Herman, J., & Yeh, J. Early childhood education. Los Angeles, CA: Center for the Study of Evaluation, 1978.

Baker, E. L., & Quellmalz, E. S. Results of pilot studies. Center for the Study of Evaluation, University of California, Los Angeles, 1980.

Herman, J. The relationship of individualized instruction variables and second grade reading, mathematics and affective outcomes. Unpublished doctoral dissertation, University of California, Los Angeles, 1977.

Hively, W. Introduction to domain-referenced testing. Educational Technology, 1974, 14, 5-10.

Levin, J. R., & Lesgold, S. N. On pictures in prose. Educational Communication and Technology, 1978, 26, 233-243.

Minsky, M. A framework for representing knowledge. In P. H. Winston (Ed.), The Psychology of Computer Vision. New York: McGraw-Hill, 1975.

Pitts, M. The relationship of classroom instructional characteristics and writing in the descriptive/narrative mode. Report to the National Institute of Education, Los Angeles: UCLA Center for the Study of Evaluation, 1978. (Grant No. OB-NIE-G-78-0213)

Quellmalz, E. S., & Baker, E. L. Effects of alternative scoring options on the classification of entering freshmen writing competencies. Report to the National Institute of Education, Los Angeles: UCLA Center for the Study of Evaluation, 1981. (Grant No. OB-NIE-G-80-0112)

Quellmalz, E. S. Interim Report. Defining writing domains: Effects of discourse and response mode. Center for the Study of Evaluation, University of California, Los Angeles, 1979.

Quellmalz, E. S., & Capell, F. Defining writing domains: Effects of discourse and response mode. Report to the National Institute of Education, November, 1979 (Grant No. OB-NIE-G-78-0213 to the Center for the Study of Evaluation.)

Spooner-Smith, L. Investigation of writing assessment strategies. Report to the National Institute of Education, November, 1978. (Grant No. OB-NIE-G-78-0213 to the Center for the Study of Evaluation.)

Winters, L. The effects of differing response criteria on the assessment of writing competence. Report to the National Institute of Education, November, 1978. (Grant No. OB-NIE-G-78-0213 to the Center for the Study of Evaluation.)

Deliverable - November 1981

CONSTRUCT VALIDITY IN WRITING ASSESSMENT: PRACTICING WHAT WE PREACH

Annual Report
Edys Quellmalz, Project Director

Grant Number
NIE-G-80-0112
P-3

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California - Los Angeles

Construct Validity in Writing Assessment:

Practicing What We Preach
(Effects of Time and Stragegy Use on Writing Performance)

Although writing is one of the three basic skills, it has received much less attention in research, instruction, and assessment than have the other two subjects areas, reading and mathematics. Now, however, accountability and minimum competency testing mandates have begun to expose the lack of understanding and attention to writing. The urgent need among school practitioners for a reliable, economical system for assessing student writing ability has led to the measurement of writing through readily identifiable, countable, text features. Accordingly, testing research issues in writing have focused upon reliability: rating scales, rating procedures, and task parameters such as selection of topic or mode of discourse. This narrowed perspective on writing raises a second measurement issue, validity. To what extent can we feel confident that writing assessment procedures are measuring nontrivial writing skills? A growing number of researchers in writing skills voice grave doubts about the construct validity of prevalent rating scales and testing methods. For the most part, these people cite erroneous measurement assumptions that emphasize a static and decontextualized written product; others express concern for the apparent lack of theoretical basis for many measurement decisions (Emig & Parker, 1976; Gere, 1980; Hirsch, 1977; Odell & Cooper, 1980; Polin, 1980; Smith, 1979).

In contrast to practitioners and test developers, most researchers have focused upon establishing and validating theory and theory-based models of writing. The most recent, successful, and widely endorsed efforts propose a dynamic view of writing as a set of "recursive" processes. In particular, cognitive information-processing and problem-solving theories applied to writing have resulted in similar models of the writers' active engagement of the writing task (Hayes & Flower, 1979; Nold, 1979). A growing amount of process-based research supports these cognitive models (see for example, Bracewell, Bereiter, & Scardamalia, 1980; Bracewell, Scardamalia, & Bereiter, 1980; Matsuhashi & Cooper, 1978; Perl, 1979; Stallard, 1978).

Briefly, the model of writing this study assumed can be characterized as a cognitive, information-processing model, comprised of two major interdependent and overlapping processes: composition and transcription. Composition refers to the invention of the message context, to activities occurring before writing; transcription refers to the encoding of the message, the actual production and refinement of the message (Stallard, 1976). These two large processes subsume many subtasks and skills.

During composing, the writer makes decisions about the audience, writing purpose and topic. These decisions act as focusing and refining criteria which influence the recurrent search and selection activities that shape the message during writing. That is, before actually writing, the competent writers conceptualize their intentions. This framework, in turn, acts as a plan guiding writing. Such a plan, e.g., the "intended meaning representation" (Nold, 1979), may affect organization, amount and kind of detail and summary generalizations, tone and even syntax in the written product.

The transcription process also can be broken down into subtasks. These include "recursive" or recurrent planning and revising during writing, massed revision efforts, and editing, e.g., of mechanics, diction, spelling. These activities require reading, or rereading, the text during and after writing, to formulate a sense of what has been produced thus far. The writers then compare this "text meaning representation" (Nold, 1979) with their original intentions. Any resulting dissonance suggests appropriate revision strategies, carried out through deletion, substitution, addition, or rearrangement of the text (Sommers, 1979).

This model presents writing as a complex activity involving many subskills and processes, each of which draws upon an individual writer's limited resources (attention, effort) and capacities (memory), as well as stores of information about the writing topic and the reading audience. Thus, in human information-processing terms, writing can be viewed as both a resource-limited and a data-limited task.[1] The effect of resource demands from the myriad activities or subskills required for competent writing performance has been termed "writer overload" (Nold, 1979). However, although writer resources are limited, they may be stretched or augmented. For instance, the writer may become more adept at some of the subtasks, thus free to "pay less attention" to them. This concurs with descriptions of "skilled writers" at work (Hayes & Flower, 1978; Matsuhashi & Cooper, 1978; Stallard, 1974). Or, the writer may employ a "metaplan" describing strategies for efficient deployment of resources across the required subtasks (Flavell, 1976; Miller, Galanter & Pribram, 1968). This may describe the implicit goal of instruction in pre-writing activities that are often explained in such terms (Odell, 1974; Young, Becker, & Pike, 1970). Another means of stretching writers' resources is to introduce more information into the task, cueing the writers and thereby reducing the processing requirements for attention. In effect, the writing task procedures may be manipulated to assume some of the burden of the many processes required to compose and transcribe the written response, an essay. In such a case, the resource-limited model of writing suggests writing performance ought to be facilitated, improved.

This last method, i.e., manipulation of task components, describes the methodology this study employed to examine the writing process construct and its implications for test design in writing assessment. This approach has been termed "facilitative intervention" and used in construction and validation of cognitive models of behavior and, in particular, in studies of writing instruction (Bereiter, Scardamalia, & Bracewell, 1979). This study did not, however, employ an instructional intervention and then test the sensitivity and fidelity of dependent measures. Instead, the study intervened in the assessment phase, breaking apart the usual assessment task, writing an essay on a given topic, into subtasks identified with predominant instruction of theory and with empirically supported theories of the writing process construct.

Clearly, a lot of mental activity occurs before, during, and after the writing of an essay. Cognitive theorists include these activities and their simultaneous, interdependent nature, in the description of writing skills. Given such a rich process domain, testing writing by scoring essay samples seems a questionable evaluation of writing competence or achievement. While essay writing surely calls upon the writer to perform all of these skills, it can adequately measure only the extent to which the student writer is able to "put it all together" in producing an essay. That is, there is no rating given for skill at correctly interpreting the topic, audience, or purpose of a given task. There is no rating describing competence at planning and revising skills. Nevertheless, research and theory suggests these are the basics upon which the essay is built. Furthermore, these subskills or processes are accepted by teachers. The publications of the National Council of Teachers of English and the focus of instructional methods consortia (the Bay Area Writing Project and its spin-offs in forty-one states) endorse and encourage process instruction in writing.

To the extent that the cognitive process models of writing are viable, current testing is short-changing both the student writers and their teachers by testing and "scoring" only the criterion performance, integrative essay production on a given task. There may be enroute or prerequisite skills and writing processes at which students are, in fact, competent performers and for which they have received effective instruction. Yet these students' competencies and growth in writing may be lost because of the "overload" arising under tests of writing requiring only an essay response to a given topic.

Method

Subjects and sampling

Tenth grade students (N=320) from two Los Angeles area high schools participated in the study. The two high schools and the study sample were racially mixed with respect to Asian, black, Hispanic, and white student groups. The schools, from different school districts, drew from middle to low-middle income neighborhoods. Students whose teachers rated them vulnerable to language interference problems from a non-English primary language were excluded from the sample (n=18).

Procedures

Within each classroom (n=13 classrooms), students were randomly assigned to one of six treatment variations. The independent variable in this study was strategy assistance. Strategy assistance refers to the worksheet activities given to students to assist them in carrying out processes that are hypothesized requisites for good writing. These processes have been described above as planning and revising. In this study there were two dimensions of planning and revising assistance: (a) broad level, Task Only, and (b) specific, level, Task-Response focused. Task Only worksheets asked students questions about the content, purpose, and audience of the given writing task. Task-Response worksheets asked students those questions interspersed with three additional questions about the same qualities in their own essay response (either for planning or revising). Thus, the four treatment groups were: (a) Planners receiving Task Only worksheets, (b) Planners receiving Task-Response worksheets. Two additional student groups wrote unassisted by worksheets: (e) Unassisted, Two-Day writers, (f) Unassisted, One-Day writers. These latter groups allowed comparisons of the effects of strategy assistance and of simply "extended time" for writing. Directions to both Unassisted groups did suggest they use their time wisely to plan, draft, and revise their essay.

Students remained in their intact classroom throughout the four consecutive days of the study. Individual study packets contained the distinctive task instructions and materials. Study monitors presided over the classrooms at all times during the study. Regular classroom teachers remained in the room, but were asked to distance themselves from the proceedings. For the most part, these teachers sat in the back of the room grading papers or planning assignments. The students did not have any difficulty understanding the nature of their daily tasks, the larger four-day context of those tasks, nor the fact that different students were engaged in different tasks. Study monitors did not identify any significant disruptions or confusions.

The study took place over four consecutive days in Spring. The first day was used to collect baseline samples of expository writing from all students. The second and third day students wrote on one of two different treatment essay topics, randomly assigned within treatment groups. The essay task asked for an expository essay written to an audience of peers (for the school newspaper) about the value(s) of either summer jobs or elective courses. Students used a blue pen on the first treatment day and a black pen on the second treatment day. This helped identify the focus of each day's efforts. Students in all groups were instructed to keep any drafts, notes or outlines they generated. These continued to be available to them through their individual study packets. The dictionary and thesaurus could be used if desired. On the fourth and last day, all students completed questionnaires on their perceptions of writing under "usual" conditions, and on their perceptions of instruction received over the semester in their English composition class.

Dependent measures

The dependent variable in this study was writing performance. Students' treatment essays were scored using two measures, each based upon very different assumptions about the nature of writing. These two measures were: (a) CSE Expository Writing Scale IV, (b) primary and secondary trait rubrics. The first scale is an analytic and holistic rating scale developed at the Center for the Study of Evaluation, UCLA (Quellmalz, 1980). A six category rubric, it rates students' essays as a whole (in categories of General Impression and General Competence) and by analyzing specific features (in categories of Essay Coherence, Paragraph Coherence, Supporting Detail, and Mechanics). In each of the six categories there is a six point range for describing competence; scores of three and below are considered below mastery or "criterion" for competence. The major assumption of the CSE Expository Scale IV is the existence of gereralizable features of good writing that can be identified with the domain of expository essay writing, regardless of topic, audience or context of the given essay task. The CSE scale and its assumption about writing reflect predominant essay measures and beliefs about writing assessment found in school districts and state educational agencies (cf. Spandel and Stiggins, 1980).

The second essay rating measure, primary and secondary trait scoring, is best described as a method of constructing both tasks and task-specific scoring rubrics. That is, this assessment system assumes just the opposite of the CSE and similar measures, i.e., writing performance is highly affected by context and content of the task. Unlike the CSE Expository Scale which can be used without modification for a variety of expository tasks, the primary trait scale is built from a careful analysis of the features of each unique writing assignment. To the extent that assignments differ, then too, the rubric requires alterations. Clearly a more labor-intensive method, the primary and secondary trait scale is more popularly endorsed by theorists and researchers than by practitioners. Nevertheless, its specificity has prompted use by the National Assessment of Educational Progress to track national writing achievement over time, and by CEMREL Incorporated in their attempt to develop a writing curriculum (Klaus, Lloyd-Jones, Brown, Littlefair, Mullis, Miller, & Verity, 1979; Mullis, 1976). For this study, essay topics were developed from a set of domain specifications, i.e., as a domain-referenced test might be. The scoring rubrics for primary and secondary traits were virtually identical for the two topics as they were constructed from the same set of task criteria. Essentially the primary trait assessed was the use of related support to link generalizations about "values" to specific features of "summer jobs" or "elective courses." The secondary trait assessed was the use of peer-appropriate referents to establish a specific audience, peers. The primary trait was rated on a zero to four point system; the secondary trait on a one to four system.

Baseline essays were obtained from all students on the first day of the study. Instructional history questionnaire data were also obtained and included in analyses. Teachers were asked to rate their students' writing ability on a six point scale for which each point was carefully defined. The baseline essays and teacher ratings were available for use as covariates in final data analyses.

Data Analyses

Three sets of analyses used essay scores derived from the two dependent measures. Analysis of variance provided answers to the major research questions about the effect on writing performance due to extended time for writing, strategy assistance, and timing and specificity of that assistance. Baseline essay scores and teacher ratings of student writing ability were used as covariates in analysis of covariance investigations of entry skills upon study treatments. Stepwise regression analyses used students' questionnaire data to determine the interaction of students' skills, instructional experiences, and perceptions with the study variables.

Rater reliabilities for the two measures were caluclated as generalizability coefficients. For the CSE Expository Scale IV, rater agreements on students' essays were: General Impression = .77; General Competence = .76; Essay Coherence = .64; Paragraph Coherence = .63; Supporting Detail = .69; Mechanics = .75. For the primary and secondary trait scores, reliabilities were .75 and .70, respectively.

The philosophical differences between practitioner and research perspectives on writing, presumed to underlie the two dependent measures, were borne out in the study data. Correlations between subscales of the CSE Expository Scale IV and the primary and secondary traits can be considered moderate at best.

- - - - - - - - - - - - - - - -

Insert Table 1 about here

- - - - - - - - - - - - - - - -

The Primary Trait Rubric emphasized building a relationship between generalizations bout "values" (of summer jobs or non-basic skills classes) and the particular features (of working or a given class) that facilitate or generate those values. This might be considered an emphasis upon support and coherence at the essay level. Not surprisingly, those were the CSE subscales with

Table 1

Correlations Among Dependent Measures

	Primary Trait Rubrics	
CSE Expository Scale IV	Primary Trait	Secondary Trait
General Impression	.47	-.10
General Competence	.42	-.11
Essay Coherence	.47	-.12
Paragraph Coherence	.32	-.06
Support	.49	-.13
Mechanics	.28	-.12

Note. With the exception of Paragraph Coherence, correlation coefficients are based upon a sample of 230 student essays. Paragraph Coherence figures reflect a sample of 139 student essays in which paragraphing was attempted. Essays without evidence of paragraphing by indentation or line skipping between blocks or prose, were assumed to be one-paragraph essays. To score them in Essay Coherence and Paragraph Coherence would mean to score the same skill twice. They were scored as "missing data" for Paragraph Coherence. This practice is reflected in all analyses of this report.

which the Primary Trait scores were most strongly, albeit moderately, correlated (r = .49 and .47, respectively). However, as might be expected from the underlying differences in perspectives on writing, Primary Trait and CSE Scale score correlations were only moderate, though significant (ranging from r = .32 to .49). The Mechanics subscale was a standout exception. Primary Trait is intended to "overlook" mechanics and syntax errors in favor of assessing students' fulfillment of the communicative intent of the task. Thus, the lower correlation, r = .28, was not unexpected. The Secondary Trait rubric emphasized audience sensitivity in students' essays. This sensitivity was defined in terms of tone, wording, and content referent markers throughout the text. As the CSE Expository Scale IV attended to audience concerns only as part of its General Impression rating, the lack of significant correlation between Secondary Trait and CSE subscale scores was expected (coefficients range from r = -.06 to -.13). It is worth noting that all correlations between Secondary Trait and CSE subscales were negative, though low.

Correlations among the subscales of the CSE Expository Scale were quite high, a phenomenon observed in some previous studies using earlier versions of the scale (Quellmalz & Capell, 1980).

- - - - - - - - - - - - - -

Insert Table 2 about here

- - - - - - - - - - - - - -

In particular, General Impression, General Competence, Essay Coherence and Supporting Detail ranged from .87 to .96. Paragraph Coherence, although less strongly correlated with other subscales, was still quite high with values ranging from .80 to .92. Mechanics was the most independent of the six categories, as might be expected. Nevertheless, moderate, significant coefficients were obtained for correlations with other subscales (ranging

Table 2

Correlations Among Subscales of the CSE Expository Scale IV

	General Comptence	Essay Coherence	Paragraph Coherence	Support	Mechanics
General Impression	.92	.93	.77	.96	.65
General Competence	--	.87	.72	.89	.63
Essay Coherence		--	.78	.90	.58
Paragraph Coherence			--	.80	.45
Support				—	.53
Mechanics					--

Note. With the exception of Paragraph Coherence, correlation coefficients are based upon a sample of 223 student essays. Paragraph Coherence reflects a sample of 139 student essays in which paragraphing was attempted and could therefore be judged.

from r = .45 to .65). In regression and some post hoc analyses, described below, the four, highly correlated subscales were collapsed (by straight averaging) into a single CSE Expository Text Level Score; Mechanics remained intact.

The student questionnaire data were selectively used to create seven instructional variables describing the instructional emphases and practice opportunities in students' English composition instruction during the semester. These variables were: practice with extended time available for writing; instruction and practice on planning; instruction and practice on revising; instruction on organization and support; instruction and practice on audience considerations in writing; instruction on grammar and punctuation; and, practice on expository writing. In addition to these instructional variables, regression analyses used students' self-reported use of revision and planning strategies, and interaction terms suggested by the correlation matrices for these variables.

Results

The variables in the study included: time available for writing (one versus two day time period); strategy assistance (worksheets or no assistance); timing of assistance (as planning or revising); specificity of assistance (three questions about the assignment versus six questions about the assignment and the essay response). The study investigated the cognitive, information-processing model of writing which describes the writing process construct in terms of numerous interdependent skills and subtasks involved in generating a competent essay. Theory postulates that this complex of processes may overwhelm writers and inhibit their performance of individual skills at which they may be competent. The study sought to alleviate this "writer overload" effect by breaking apart the writing task to allow students to focus their efforts upon planning and revising process requirements. This was expected to facilitate performance by allowing students' skills, e.g., at planning and revising, to be fully realized.

Extended Time for Writing

To determine the effects of strategy assistance on writing performance, it was necessary to be able to exclude the effect of simply having more time available in which to write. For that reason, the two-day, Unassisted Group was included. It was also valuable to know whether any increment in performance in the two-day, Unassisted Group would arise; for this reason, the study also included the traditional, writing assessment setting: one day for one essay.

The comparison between the two day and one day, Unassisted student writers indicated that there was no statistical difference in their performance. This result held true for both dependent measures in both the analysis of variance and covariance procedures.

The student questionnaire asked students about the significance of time constraints in a variety of contexts. First, asked about preferences for writing assessment, 69.8% of the students (n = 208) indicated they would rather write one essay over two days, than write two essays in the standard time frame of one class period for each (usually about fifty minutes, less "settling down" time). Second, time constraint on writing performance was raised as a possible problem student writers faced, among other possibilities such as worrying about "what the teacher wants," about "good grammar," what to say in the essay, and so forth. Table 3 presents data on this second question about time restrictions.

- - - - - - - - - - - - -

Insert Table 3 about here

- - - - - - - - - - - - -

Table 3

Students' Self-Reported Writing Problems

Writing Problems	Students Indicating Yes Number	 % of Total[a]
Coming up with ideas to use in the essay	145	49.8
Organizing my ideas	123	42.3
Finishing before I run out of time	122	41.9
Figuring out what the teacher wants	113	38.8
Getting down on paper the ideas I have in my head	108	37.1
Writing in "good" English	106	36.4
Knowing what to do to make the essay better	95	32.6
Going back to check over what I've written	40	13.7
Nothing, I don't have much trouble writing essays.	33	11.3

Note. Students checked multiple responses; column total exceeds 100%.

[a] The number of students answering the question totalled 291.

"Finishing before I run out of time" ranked third among the list of nine potential concerns for student writers. Forty-one percent of the students (n = 122) indicated time constraint pressure was a problem for them. The third context in which students were queried about time was in terms of behaviors and processes going on during actual writing. Table 4 presents results for this question.

- - - - - - - - - - - - - -

Insert Table 4 about here

- - - - - - - - - - - - - -

Included along with "thinking about how much time is left" were such choices as rereading, editing, planning, rethinking ideas. In this grouping, time concerns were much less salient than other concerns, and ranked fifth on a list of eight items. However, 41% (n = 122) again indicated watching the clock was something they were conscious of doing while writing.

The student questionnaire also inquired about students' experiences with time limits on writing. Table 5 presents these data.

- - - - - - - - - - - - - -

Insert Table 5 about here

- - - - - - - - - - - - - -

Here a majority of students, 58.9% (n = 175), indicated that they had had four or more opportunities during the semester to write their essays over a longer period of time than one class session or one "overnight" period. Only 16.5% of the respondents reported never or only once writing under such extended time conditions.

In short, students report time constraints are a major concern during writing, as a general writing problem, and in consideration of writing assessment preferences. It appears that a majority of students have regular opportunities to write essays without time constraint pressures. Nevertheless, despite their expressed preference for and experience with extended time

Table 4

Students' Self-Reported Processes During Writing

	Students Indicating Yes	
Processes During Writing	Number	% of Total
Planning ahead for the next thing to say	199	68.4%
Rereading what I've written, even before I'm finished	190	65.3
Changing my mind about ideas	179	61.5
Fixing spelling, grammar and/or punctuation mistakes	154	52.9
Thinking about how much time is left	122	41.9
Trying to keep in mind who's going to read the essay	56	19.2
Starting over lots of times	37	12.7
Unsure (or) I just keep writing until I'm through	28	9.6

Note. Students checked multiple responses; column totals exceed 100%.

[a] The number of students answering the question totalled 291.

Table 5

Students' Reports of Opportunity to Practice

	Number of Occasions in the Semester			
Practice on Instruction	None	Once	Two to Three	Four or More
Spend more than one class period or one night writing an essay	8.4%	8.1%	24.6%	58.9%
Write an essay as if someone besides the teacher were going to read it	37.9	17.4	17.4	28.2
Write essays in other classes like history or science	46.1	14.2	18.0	21.7
Turn an essay back into the teacher, after you rewrote all or part of it	28.5	17.1	22.1	32.2
Turn in a rewritten paper and get it back, graded, a second time	40.4	17.8	18.2	23.6

Note. Total number of respondents is 298. Figures represent the percent of total respondents indicating a particular frequency for each item.

for writing, simply providing students with that extra time did not improve their writing performance in comparison to students writing under the constrained time condition.

Strategy Assistance Effects

At its broadest level, the study investigated the impact of strategy assistance as opposed to none, i.e., the extra time only, described just above. This contrasted the Unassisted two-day writers against the assisted groups of planners and revisers. Table 6 presents the means and standard deviations for each group.

Insert Table 6 about here

The analysis of variance investigations did not yield strong effects for assisted groups on any of the CSE Expository subscales. However, when ability covariates were used, allowing greater control over the within group variation, marginal effects emerged for the Essay Coherence subscale of the CSE Expository Scale IV (p = .07). Means among strategy assisted and unassisted groups was largely due to context of that assistance.

The scores for assisted and unassisted groups looked promising for the Secondary Trait (see Table 6). It seemed that simply providing students with worksheets that focused some attention on the audience feature of the writing assignment, resulted in imporved scores on that trait. This difference did not yield stong significant results in analysis of variance (p = .11). It did seem that results might be more favorable when the context variable governing assistance was examined.

The student questionnaire also asked students whether they considered themselves planners and revisers when they normally wrote essays. The possibility that the treatment worksheets interfered or interacted with these students' planning and revising processes was considered in regression

Table 6

Means, Standard Deviations for Strategy by Specificity

		Planners		Revisers	
Subscale	Unassisted	3 Q's	6 Q's	3 Q's	6 Q's
CSE Expository Scale IV					
General Impression	3.24 (1.03)	3.19 (1.14)	3.22 (1.05)	3.20 (1.26)	3.17 (1.10)
General Competence	3.05 (0.95)	3.24 (0.99)	3.31 (0.87)	3.15 (1.08)	3.19 (1.06)
Essay Coherence	3.21 (1.00)	3.58 (0.99)	3.54 (0.91)	3.38 (1.12)	3.33 (1.13)
Paragraph Coherence[a]	3.30 (0.99)	3.43 (0.95)	3.43 (0.81)	3.46 (1.05)	3.38 (0.94)
Support	3.22 (1.03)	3.29 (1.04)	3.45 (0.83)	3.38 (1.21)	3.40 (1.04)
Mechanics	3.55 (0.99)	3.51 (0.99)	3.59 (0.88)	3.29 (0.91)	3.24 (0.97)
Primary Trait Scoring Rubrics					
Primary Trait	1.14 (0.69)	1.13 (0.63)	1.05 (0.75)	1.16 (0.80)	1.28 (0.73)
Secondary Trait	1.24 (0.72)	1.41 (0.66)	1.39 (0.70)	1.59 (0.87)	1.52 (0.64)
Number	43	55	39	41	40

[a] The Paragraph Coherence subscale is scored only where paragraphing has been attempted. Non-paragraphed essays or single paragraph essays were considered cases of missing data for the subscale. Accordingly, the group size differs for that subscale, as noted: Unassisted n=28; 3 Q's Planners n = 34; 6 Q's Planners n=30; 3 Q's Revisers n=25; 6 Q's Revisers n=24.

analyses entering self-reported strategy as a variable predicting dependent measure scores on the CSE Scale composite text-level scores (described earlier), the Mechanics subscale score, the Primary and the Secondary Trait scores. These results are displayed on Tables 7a through 10b, and are discussed in the next section on context and specificity of worksheet assistance, with which they did, in fact, interact.

- - - - - - - - - - - - - - - - -

Insert Tables 7a - 10b about here

- - - - - - - - - - - - - - - - -

Context of Strategy Assistance

The two values for the context of strategy assistance describe the timing of worksheet assistance in relation to the two-day writing process. Students who received worksheets at the beginning of the first day were encouraged and assumed to apply that assistance in planning their essay. Students who received their worksheets at the beginning of the second day were assumed to apply that assistance in revising their essay. Students were obliged to compare their worksheets before moving on to any writing or rewriting activities. The colored worksheets were easy to spot in the classroom and monitors for the study were able to ensure that students did complete their worksheets as scheduled. Students marked down the starting and finishing times for completing the worksheets. The average time for the six question worksheets was twenty-five minutes; fifteen for the shorter worksheets. There did not appear to be a difference in time between Planning and Revising contexts.

Analysis of variance did not reveal any strong differences between Planning and Revising groups on the CSE Expository Scale. However, the CSE Mechanics subscale scores were significantly different for the two strategy context groups ($p = .05$). Means for the two groups suggested that, for the

Table 7a

Mechanics Scores Regressed on Strategy Treatment, Instructional Practice and Normal Strategy, and Interactions

Source	b	Beta	f ratio
Strategy Treatment, Step 1			
Plan	-.08	-.04	.22
Revise	-.26	-.13	2.26
Practice and Normal Strategy, Step 2			
Extended Time for Writing	.17	.25	9.81*
Practice with Audience, Tone	-.19	-.21	6.75*
Practice with Expository Mode	.13	.18	5.58**
Interactions, Step 3			
Revise Treatment x Normally Revise Only	-.69	-.15	5.09**

*p _ .01

**p _ .05

Table 7b

Correlations for Variables in the Regression Equation:

CSE Mechanics Subscale Scores

Variables	1	2	3	4
1. Extended Time for writing	—	.54	.45	.04
2. Audience Practice		—	.41	.18
3. Exposition Practice			—	.01
4. Usually Revise Only x Revising Assistance Group				—

Note. The variables in the table are instructional variables created from the questionnaire data.

Table 8a

Expository Scale Regressed on Strategy Treatment,
Instructional Practice and Normal Strategy, and Interactions

Source	b	Beta	F
Strategy Treatment, Step 1			
Plan	1.47	.07	.83
Revise	-.04	.00	.00
Practice and Normal Strategy, Step 2			
Extended Time for Writing	3.35	.47	36.46*
Practice on Revising Activities	-2.78	-.33	13.20*
Practice with Audience, Tone	-2.12	-.23	5.93**
Practice with Expository Mode	.88	.11	2.60
Interactions, Step 3			
Normally Plan and Revise x Practice on Organization and Support	1.30	.16	5.40**
Normally Plan Only x Practice on Revising Activities	2.20	.12	3.58

Note. Analytic Score represents average over CSE Expository Subscales: General Impression, General Competence, Essay Coherence, Supporting Detail. Refer to Results section for explanation of conversion.

*p _ .01

**p _ .05

Table 8b

Correlations for Variables in the Regression Equation:

CSE Expository Scale Score[a]

Variables	1	2	3	4	5	6
1. Extended Time	--	.52	.54	.45	.35	.03
2. Revising Practice		—	.73	.33	.41	.02
3. Audience Practice			—	.41	.31	.10
4. Exposition Practice				—	.24	.11
5. Usually Plan & Revise x Organization & Support Practice					--	-.25
6. Usually Plan Only x Revising Practice						--

Note. The variables in the table are instructional variables created from the questionnaire data.

[a] The CSE Expository Scale score is a composite score representing the average over the subscales of General Impression, General Competence, Essay Coherence, and Support. This transformation is described in the text of the Results section.

Table 9a

Primary Trait Score Regressed on Strategy Treatment, Instructional Practice and Normal Strategy, and Interactions

Source	b	Beta	F
Strategy Treatment, Step 1			
Plan	.01	.01	.01
Revise	.08	.06	.42
Practice and Normal Strategy, Step 2			
Normally Revise Only	-.23	-.12	-2.35
Normally Plan and Revise	.19	.14	2.96
Extended Time for Writing	.08	.16	3.34
Practice on Revising Activities	-.16	-.26	10.84*
Practice with Expository Mode	.05	.04	1.66

*p _ .01

Table 9b

Correlations for Variables in the Regression Equation:

Primary Trait Rubric Score

Variables	1	2	3	4	5
1. Usually Revise Only	--	--	.12	.23	.04
2. Usually Plan & Revise		—	.37	.21	.19
3. Extended Time for Writing			--	.52	.45
4. Revising Practice				--	.33
5. Exposition Practice					—

Note. Variables in the table are from the questionnaire responses.

Table 10a

Secondary Trait Scores Regressed on Strategy Treatment, Instructional Practice and Normal Strategy, and Interactions

Source	b	Beta	F
Strategy Treatment, Step 1			
Plan	.21	.15	.13
Revise	.32	.22	6.01**
Practice and Normal Strategy, Step 2			
Practice on Planning Activities	-.14	-.21	6.81*
Practice on Revising Activities	.23	.35	20.83*
Interactions, Step 3			
Revise Treatment x Normally Plan Only	-.35	-.13	3.72

*p _ .01

**p _ .05

Table 10b

Correlations for Variables in the Regression Equation:

Secondary Trait Rubric Score

Variables	1	2	3
1. Planning Practice	--	.53	.12
2. Revising Practice		—	.45
3. Usually Plan Only x Revising Assistance Group			—

Note. Variables in the table from the questionnaire responses.

Mechanics score, the Revising groups were degrading their scores, rather than that the Planning group students were somehow performing better on this dimension. Note that the Mechanics mean for the Unassisted Group is 3.55, and for the Planners, 3.54 (see Table 6 for means). Post hoc comparisons using Scheffe's test for significance supported this hypothesis ($p = .04$).

The two assistance groups also differed significantly in their Primary Trait scores ($p = .06$), when teacher ratings were entered as a covariate to account for additional within-group variation. Means for the Planners and Revisers revealed that the Revisers were outperforming the Planners (see Table 6). It appeared that this difference was primarily the result of better scores for the Revising Group with the six-question worksheet assistance. That is, the effectiveness of assistance was tempered by the interaction of context and specificity for that assistance.

The Secondary Trait score differences for the Planning and Revising looked more promising than the Primary Trait scores had. However, the comparison of assisted (Planning and Revising combined) and Unassisted groups had not turned up statistically significant differences, despite the apparent difference in means. Differences between Planners and Revisers on this score were also nonsignificant ($p = .13$).

Specificity of Strategy Assistance

Specificity of assistance describes the distinction between the long, six-question worksheet and the shorter, three-question worksheet used by Planners and Revisers. The short worksheet asked students to decode the essay assignment in terms of its audience, content, and purpose. The longer versions also asked students to either develop a plan in response to those features, or to interpret their own essay draft in light of these features of the assignment.

Contrasting the three and six question writers (across the planning and revising context), analysis of variance and covariance did not yield any differences for this main effect of specificity of assistance. However, interactions of specificity and context for strategy assistance did yield some interesting results.

Interaction of Strategy Context and Specificity of Assistance

Means for all four strategy context by specificity groups suggested a few comparisons might reveal interaction effects for these variables. First, mean scores on the CSE Support subscale appeared to differ for the two Planning groups, depending upon which version of the worksheet student writers were exposed to. The short version, Three-Question Planners, averaged 3.29, compared to the Unassisted Group which averaged 3.22. The Six-Question Planners, however, averaged 3.45 on the same subscale. This mean was the highest of all group means. The Revising groups, both the Three- and Six-Question Revisers, did not appear to be that much different in their scores. Analysis of variance did not turn up any interaction effects for the CSE Expository Scales, including the Support subscale. However, when the covariates were used, the interaction of specificity and context did attain some significance ($p = .05$). Post hoc comparisons using Scheffe revealed the expected marginal difference between the Three- and Six-Question Planning groups ($p = .06$).

Although the Primary Trait means looked promising for the effects of Six-Question worksheets by Planning versus Revising context, this difference did not test out at a significant level under Scheffe ($p = .16$).

When the four highly correlated subscales of the CSE Expository Scale were collapsed into a single "composite" Text-level score for exposition, the context by specificity interaction effect was marginally significant for the Planning groups in analysis of variance ($p = .06$). The Six-Question Planners outscored the Three-Question Planners.

Regression Analyses with Instructional Variables

For the composite CSE Expository Text-level Scale score, instructional practice that allowed extended time for writing a single composition yielded significantly higher scores (p = .01). Other instructional variables of significance bore a negative relationship to the Expository Scale Text-level score. Instruction and practice in revision showed a strong, negative relationship with essay scores (p = .01). Instruction and practice on audience concerns in writing, also, though less strongly, demonstrated a detrimental influence on the Expository Scale (p = .05). Students who reported themselves as both "planners and revisers" and who reported greater instructional emphases on organization and supporting detail in writing, scored more highly on their essays (p = .05).

On the Mechanics subscale, higher scores were found for students reporting greater practice in expository writing (p = .01), and essay writing that extended beyond one class period (p = .05). Interestingly, negative influences on scores were found for students reporting greater instruction and practice with audiences besides the teacher/evaluator and audience considerations such as style and tone (p = .01). Students in the Revising treatment who reported themselves to be "revisers only," had significantly lower Mechanics scores (p = .05).

For Primary Trait scores, the only significant variable was the negative influence from instruction and practice on revision (p = .01). The Secondary Trait scores obtained the opposite result; higher scores for students reporting greater instruction and practice on revision (p = .01). Revision treatment group membership also resulted in higher Secondary Trait scores (p = .05). However, lower scores resulted for students reporting more instruction and practice emphasizing planning activities (p = .01).

Discussion

Summary of Results

This study was successful in its attempt to break apart the essay test task into meaningful subtasks. The domain of writing skills, defined to include a cognitive process construct, is indeed legitimate to the extent that the study was able to operationalize some of those processes for students.

Interestingly, treatments interacted with the two philosophically distinct measures. Planning-assisted students were superior to other students on the analytic scale categories, except mechanics. Revising students degraded their scores on the mechanics scale. On the other hand, primary trait scores were higher for revising students; for the secondary trait, essay audience, all assisted students outdid unassisted peers.

Regression analyses using questionnaire data confirmed the study premise. Students who reported themselves as "planners only" were immune to the negative effects (on the mechanics scale) in prompted revision groups; students who reported that they were "revisers only" had their revision problems exacerbated by encouragement to revise. Students reporting themselves as both "planners and revisers" were more effective regardless of the treatment group in which they found themselves.

Interpretation of Results

Strategy assistance treatment made a difference in the subsequent writing performance of a significant number of students in both the primary/secondary trait scale and the CSE analytic scale. In General Essay Coherence ratings, students who first completed planning sheets scored higher than their revising (and unstructured) peers. In the Support scale, this effect of planning depended upon the level of specificity of our prompting. Students planning with the six-question worksheet scored well above other worksheet prompted groups. On the primary trait subscale, strategy was again a potent effect. However, here the revising sheet students outscored the planners. Planning effects then, seem more salient when assessment methods emphasize test qualities presumed to exist in all writing, i.e., generalized writing skills. On the other hand, revising seems more effective in assessments that emphasize communicative intent, i.e., skill in addressing the purpose and audience of the writing, although it is unclear why this effect would not have been mirrored by the General Impression score of the analytic scale.

Planning worksheets had students decode the given task in terms of the content, the purpose and the given audience. It is, of course, true that the impact of answering such items may be effective merely because it slows students down to read the topic carefully. However, the differential effectiveness of the specificity of planning activities suggests that something more was going on, at least for the detailed planners. Further interpretation of results should note differences between the analytic and primary trait scale. Decision rules for classifying a paper "off topic" differed considerably. Frequently, otherwise "well written" (in a text sense) essays were included and rated highly by CSE analytic scale raters, whereas the same essays were judged "off topic" by primary trait raters. Further, the analytic scale completely eliminated "off topic" essays, scoring them as cases of "missing data," while the primary trait scale relegated off topic essays to the lowest category of competence.

Also, the analytic scale assumed that its six categories measured separable text features, with the General Impression and General Competence scores functioning as global, composite judgments (see Results section). However, the correlation between even the four presumably discrete text elements, Essay Coherence, Paragraph Coherence, Support and Mechanics were very high, with the exception of the Mechanics scale. Therefore, when examined across all treatment groups the subscale intercorrelations suggest that the CSE scale provided two distinct scores, one at the text-level, one at the sentence level. In effect, the CSE scale, with its five related scores, might have been able to account for more variation in raters, i.e., to be more sensitive to effects.

These measurement factors aside, writing theory offers suggestions regarding findings. The planning students might in fact have been led to formulate a representation of the task features before beginning to write. This sense of task, theory proposes, guided writers in drafting their essay response. Thus we expected such guidance to be reflected in an "essay coherence" subscale. That is, planners had formulated a task context (parameters) within which to write. Further, the specific level planners who were prompted to plan their own essay, had rehearsed possible content before writing. Our model suggests that by relieving some of this "thinking while writing" load, we should have facilitated performance. For the specific planners, improved writing performance was reflected, these students, additional planning questions asked students to plan for main idea and supporting details, as well as audience factors to consider. It is gratifying to have the effects of structured planning show up as significantly higher scores on the supporting detail subscale.

Revising effects were confined to the primary trait scale, and the impact was uneffected by specificity of revising activities. Revising worksheets also asked students to decode the task (again, in terms of content, purpose, and audience). It is true that revising effects might simply have resulted from a "break" from writing and a fresh return to the task after completing revising worksheets. However this "break" effect was also available to planners, most of whom began drafting their essay on the first day, returning to finish it on the second day. That differences in the specificity of revision worksheets did not make a difference suggests that the general task decoding questions were sufficient. Simply the reconsideration of the given task appears to have provided revising students with some input to improve their essays.

Why did revising affect only primary trait, and planning only the analytic scale? Two important assumptions in each of the measures may account for effects. In primary trait scoring, raters were cautioned specifically against letting students' grammar and mechanics interfere with judgments of the primary trait. For both essay topics, the primary trait scale emphasized the writer's ability to build a relationship between specific reasons and a general resolution. Secondly, this scale's highly task sensitive categories resulted in a very clear distinction between "off topic" and "eligible" essays. Under the CSE analytic scale rubric, few essays were deemed "off topic."

Planning, however, did not include a "check" on the validity of the student planners' interpretation of the essay task. That is, planners might initially go "off topic" and without a critical reappraisal of match between task and essay response, never realign their essay. Unprompted students tend to constrain their revision efforts to word and sentence level modification. Planners, then, judged under a scale emphasizing task and essay match, and de-emphasizing text features, might lose their advantage against a group of revising students and to larger, task-oriented reconsiderations.

Revisers began writing without planning assistance. Obviously, "off topic" responses were also likely to be generated (perhaps even more likely so). However, revisers were halted somewhere between drafting and turning in the essay to be rated. In this writing hiatus, students were asked to go back to the original task or assignment and decode it in terms of its content, purpose and audience. After this "time out" to reconsider the given task, revisers returned to their essay responses, the representation of the task components fresh in their mind. Further, these students had been cued to revise their essays in light of their worksheet responses. That is, they had been prompted to view revision in a broader context. Thus for revisers, text level features such as the use of transitions (valued in the essay coherence subscale) were less salient than the alignment of task and response. Accordingly, it is not surprising that primary trait scores were improved by revising activities. The absence of this effect from the analytic scale seems a bit difficult to explain, except in terms of the comparative strength of planning effects. It may be less that revising is ineffective for the analytic scales, than that the planning effect is simply greater. In fact, if we look at strategy group means, revisers do outperform the comparison group (two-day, unstructured) but nevertheless trait planning group effects.

In and of itself, students' usual writing strategies (self-reported) affected scores on the primary trait scale when students were assisted by planning prompts, and particularly when this assistance was more detailed. Students who reported naturally employing planning strategies were more successful with their essays. The earlier reported effects were stonger. Additionally, previously unaffected subscales of mechanics (CSE analytic scale) and audience on the secondary trait subscale reflected the impact of strategy treatments when ability (as usual strategy use) was entered into the model.

Thus it appears that students may indeed be able to use strategies, planning at least, yet not be able to bring their strategy skills to bear upon their essay. This is less the case for revision; however, instructional history, as reported by students, suggests students receive little practice or feedback on their revision efforts. While teachers correct and hand back papers with useful commentary, these remarks reflect upon the planning and writing more than the revision. They provide information about writing efforts so far, i.e., pre-revision, and supply information to guide next efforts on a new (next) assignment.

Further, post hoc revision is much more difficult to prompt. Students' planning efforts are always put into action, even subconsciously, once the writing begins. However, simply encouraging revision, even having students reread and critique essay and task (as in our revision worksheets), does not ensure that students will use their revision information and ideas, nor know how to do so. In short, treatment in revision was much less controlled by the structured context than by student cooperation and effort.

Implications for Test Design

This study succeeded in helping students divide the essay task into subtasks they could handle. This expanded domain of writing skills included planning and revising processes and without supplying answers, asked students to focus on the main features of the essay assignment. In particular, success was greatest for planning processes. Using a worksheet with items requiring a written response, students in planning groups decoded the test task into features of audience (peers), content (topic 1 or topic 2). Presumably this supplied writers with a "representation" (albeit crude) of communicative intent and task parameters. This alone made a difference in writing as the planning students drafted their essay responses. A planning subgroup was led further into planning processes by answering worksheet items about the actual main idea and support, and features of the given peer audience that would formulate the essay response. In short, these planners made "plans to do" and

and "plans to say." Carrying out these processes by direction and over an extended period of time (not at the expense of essay writing time), prompted writers seemed better able to cope with writing demands and produce better essays.

Results from this study bear upon test design in writing. It appears that there are, in fact, subskills involved in writing that affect the quality of writing whether that quality is defined in terms of text features or communicative function. It appears that these subskills or processes can be broken out from the essay writing task. It also appears that many students who claim to perform these subskills are unable to do so effectively in their essay writing. This suggests that if we only measure writing in terms of the complex, integrative task of generating a complete essay we are missing student competence at lower levels of skill (developmentally or hierarchically).

On the other hand, if we expand assessment of writing to sample the full domain of writing skills we may provide more instructionally useful and sensitive information about student competence. Perhaps we should explore the possible methodologies for assessing "enroute" skills such as planning and revising (beyond simple word and sentence level errors), even determining communicative purpose of writing.

A second important implication of this study is the measurement focus issue. The differential emphases of the analytic and primary trait scale were reflected to some extent in the difference between strategy foci. Where planning led to greater cohesion and support, revising led to greater success at fulfilling the task purpose and attending to audience considerations. Although it is unclear why the primary traits coordinate emphases on supporting generalizations with specific detail did not correlate more strongly with the analytic scale ratings of coherence and support. Further research efforts might attempt to disentangle concerns for audience and task sensitivity by comparing separate ratings of these features with the composite primary trait rating.

In sum, we believe this study has provided rationale, empirical support and some avenues to explore in the development of a broader, more valid assessment approach to the writing skills area.

Reference Notes

1. Bracewell, R., Bereiter, C., & Scardamalia, M. How beginning writers succeed and fail in making written arguments more convincing. Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.
2. Bracewell, R., Scardamalia, M., & Bereiter, C. An applied cognitive-developmental approach to writing. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.
3. Bracewell, R., Scardamalia, M., & Bereiter, C. A test of two myths about revision. Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.
4. Emig, J., & Parker, R. Responding to student writing: Building a theory of the evaluating process. Rutgers University, 1977.
5. Hayes, J. R., & Flower, L. Protocol analysis of the writing process. Paper presented at the annual meeting of the American Educational Research Association, Tornoto, 1978.
6. Hayes, J. R., & Flower, L. Writing as problem solving. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.
7. Matsuhashi, A., & Cooper, C. A video time-monitored observational study: The transcribing behavior and composing process of a competent high school writer. State University of New York, Buffalo, 1978.
8. Mullis, I. The primary trait system for scoring writing tasks. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979 (ERIC Document Reporduction Service No. ED 124 942).

9. Nold, E. The Writing Process. Unpublished manuscript, Stanford University, 1979.

10. Polin, L. G. Alternative conceptions of the writing skills domain: Problems for the practitioners. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, 1980.

11. Smith, L. S. The Effects of Differing Response Criteria on Assessment of Writing Competence. Unpublished doctoral dissertation, University of California, Los Angeles, 1979.

12. Sommers, N. Revision strategies of student writers and experienced writers. Paper presented at the annual meeting of the National Council of Teachers of English, San Francisco, 1978.

Reference List

Flavell, J. H. Metacognitive aspects of problem solving. In L. Resnick (Ed.), The Nature of Intelligence. Hillsdale, New Jersey: Lawrence Earlbaum Associates, 1976.

Gere, A. Written composition: Toward a theory of evaluation. College English, 1980, 42, 44-58.

Hirsch, Ed. D. The Philosophy of Composition. Chicago: University of Chicago Press, 1977.

Klaus, C. H., Lloyd-Jones, R., Brown, R., Littlefair, W., Mullis, I., Miller, D., & Verity, D. Composing Childhood Experience: An Approach to Writing and Learning in the Elementary Grades. St. Louis: CEMREL, Incorporated, 1979.

Miller, G. A., Galanter, E., & Pribram, K. H. The formulation of plans. In. P. C. Watson & P. N. Johnson-Laird (Eds.), Thinking and Reasoning. Baltimore: Penguin Books, Incorporated, 1968.

Odell, L. Measuring the effect of instruction in pre-writing. Research in the Teaching of Writing, 1974, 8, 220-240.

Odell, L., & Cooper, C. Procedures for evaluating writing: Assumptions and needed research. College English, 1980, 42, 35-44.

Perl, S. The composing process of unskilled college writers. Research in the Teaching of Writing, 1979, 13, 317-336.

Quellmalz, E. Final Report: Controlling Rater Drift. Los Angeles: Center for the Study of Evaluation, UCLA, 1980.

Quellmalz, E., & Capell, F. Final Report: Defining Writing Domains: Effects of Discourse and Response Mode. Los Angeles: Center for the Study of Evaluation, UCLA, 1979.

Spandel, V., & Stiggens, R. J. Direct Measures of Writing Skill. Portland, Oregon: Northwest Regional Educational Laboratory, Clearinghouse for Applied Performance Testing, 1980.

Stallard, C. K. An analysis of the writing behavior of good student writers. Research in the Teaching of English, 1974, 8, 206-218.

Stallard, C. K. Composing: A Cognitive process theory of writing, College Composition and Communication, 1976, 27, 181-184.

Young. R., Becker, A., & Pike, K. Rhetoric: Discovery and Change. New York: Harcourt, Brace & World, Incorporated, 1970.

DESIGNING WRITING ASSESSMENTS: BALANCING FAIRNESS, UTILITY, AND COST

Edys S. Quellmalz

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, CA 90024

Grant No. OB-NIE-G-78-0213

The research reported herein was supported in whole or in part by a grant to the Center for the Study of Evaluation from the National Institute of Education, U. S. Department of Education. However, the opinions and findings expressed here do not necessarily reflect the position or policy of NIE and no official NIE endorsement should be inferred.

Designing Writing Assessments: Balancing Fairness, Utility, and Cost

To attain the fundamental goal of language competence, educators, students, and parents must have information describing the status and progress of language skills development. Mounting concern for student achievement in writing, one of the principal arenas of language development, has refocused the attention of policy makers, evaluators, instructors, and researchers on the features of writing assessments necessary to represent a student's writing skill fairly, usefully, and economically. While the relationship between procedures employed to evaluate writing in large scale testing and those used in the classroom has historically been tenuous, the requirements of minimum competency testing programs have stimulated research on methods to tighten the connection. These competency testing programs require school systems to assess the status of basic skill achievement, and then either to certify that minimal competencies have been attained or signal the need for remediation and provide repeated opportunities to pass comparable test forms. If these writing competency tests are to fulfil their intended function, then the writing assignments and evaluative criteria of large scale tests and classroom instruction must interrelate.

At present, many large scale writing tests bear little resemblance to students' classroom writing experiences. Many states and districts rely on multiple choice tests that measure sentence-level editing skills or passage comprehension. When writing samples are collected, the structure and topic of the writing assignment may call for information and strategies that vary considerably from students' experiences in and out of the classroom. Furthermore, writing samples are often scored rapidly and holistically by raters trained to varying levels of precision and accuracy. Students receive a single score purportedly representing the level of their writing competence.

Reactions of practitioners and researchers to such current practices are increasingly critical. They find many faults in current writing tests -- their logical and psychological relevance to realistic writing situations, their utility for informing decisions about individual competence or program effectiveness, their fairness to students and instruction, their legality for sanctioning exit requirements. This paper suggests that state and district writing assessments should re-evaluate their current methods for assessing student writing competence in light of these criticisms. An accumulating body of literature indites many of the methods assessments now use that have been derived from custom, folklore, and adaptations of norm-referenced testing methodology that are inappropriate for the purposes of competency assessment. By examining the criticisms leveled at writing tests and considering alternatives proposed by recent writing theory and research, we may find solutions that will improve the fairness and utility of writing assessments, yet remain within reasonable economic bounds.

Problem 1: Specifying Writing Goals

Just what is "good" writing? For schools, a major conflict has been to distinguish between realistic characteristics of minimum competence, reasonable high school writing exit competence and the competence of professional writers and "experts." A significant component in this controversy over "standards" has been the function various types of writing can and/or should have for the student. Thus the discourse aim or writing purpose of transactional writing has been identified by many school systems as functionally most relevant to the majority of students. At the lower grades, expressive writing has been viewed by some as valuable in its own right and by others as an educational vehicle for motivating writing that will increase fluency and sentence-level competence.

Clearly, the schools' definition of the target constrains the specific criteria that will provide logical and empirical evidence that the target has been hit. Currently, goals may relate to two competency levels, a minimum competency level targeted by most state and district minimum competency testing programs and a reasonably desirable high school exit competency level implied in many systems' curricular goals. Most competency programs emphasize transactional writing in the factual narrative, expository or persuasive modes. Minimum program goals are often that students write a clear, coherent paragraph that makes a point and that exhibits few or no mechanical, sentence-level errors. For high school exit goals, English departments set their sights at the multi-paragraph, essay level, seeking writing that has a theme or point, that is coherent between, as well as within paragraphs, and that exhibits few sentence-level errors. While minimum goals generally specify functional writing, high school desirable exit goals may expand the types of writing aims or purposes in which they would like students to be competent. By distinguishing between minimum and desirable goals, school systems may be in a better position to defend the logic, utility and fairness of focused test procedures.

Problem 2: Designing Appropriate Writing Tasks

Perhaps the most common controversy in the design of writing tests swirls about the relative merits of direct and indirect tasks. Indirect, usually multiple choice, measures have been defended by test publishers because of their economy and high correlations with essay scores (Godshalk, Swineford & Coffman, 1966; Breland & Braucher, 1977). Critics of multiple choice tests reject them logically and psychologically. They argue that multiple choice tests present primarily editing tasks or comprehension tasks and that they therefore do not tap the same kinds of mental processes required by production tasks (Bourne, 1966; Quellmalz, 1978; Cooper, 1979). Recent empirical studies of student's scores on direct and indirect measures indicate considerably lower correlations between writing skill component scores derived from multiple choice and writing samples (Quellmalz & Capell, 1979; Quellmalz, Smith, Winters & Baker, 1980; Moss, Cole & Khampaliket, in press). Furthermore, Quellmalz and Capell (1979) found multiple choice test scores provided less distinctive information about underlying writing skill constructs or traits than did essay ratings (Quellmalz & Capell, 1979). In combination, these studies support contentions that direct and indirect measures tap different psychological processes. These data would also, of course, suggest that, multiple choice test scores would not serve as fair or useful proxies for actual writing skill. At best, multiple choice tests seem to over estimate skills (NAEP, 1981) since they measure skills presumably enroute to production skills (Skinner, 1957).

In addition to the form of response required by writing tests, there is considerable disagreement about the appropriate structure of assignments used to prompt writing. Criticisms of writing tasks are that they do not present full rhetorical contexts that sufficiently inform students about the writing purpose, topic, audience, writers' role and intended criteria (Britton, 1978; Cazden, 1974; Scribner & Cole, 1978; Florio, 1979). Research shows that writers perform differentially well when writing in different discourse modes, e.g., exposition and narration (Veal & Tillman, 1971; Crowhurst, 1980; Quellmalz & Capell, 1979; Praeter & Padia, 1980; Baker & Quellmalz, 1980). Research also reveals that accessibility of information about an assigned topic affects the quality of students' writing (Baker & Quellmalz, 1980). Polin has found that when given extended time and cues about the rhetorical demands of the task during planning or revision, some writers improve some features of their work. In sum, studies of features of the writing task that influence students' writing performance suggest that variations within features such as mode of discourse (writing aim) topic, audience, time and structural cues do present different psychological demands and therefore should be distinctly specified. To be clear and fair, the writing task should provide a full rhetorical context and time to engage in all parts of the writing process. The cost of developing well formed writing prompts is not high, particularly in comparison to the cost of erroneous inferences about competence made from assessments of writing students generated in response to incomplete or ambiguous prompts.

Problem 3: Specifying Scoring Criteria and Type of Rating Scale

Criteria employed for evaluating student writing vary along a number of dimensions: from qualitative to quantitative; from general to specific; from comprehensive, full discourse features to isolated features; from vague guidelines to replicable, objective guidelines.

At the most qualitative, vague end of the continua are general impression scoring schemes where readers apply their own criteria to give the writing a single global score. Follman and Anderson's "Everyman" procedures (1967) and teachers' A-F general schemes fall in this category. Still providing a single score or quality rating, but guided by slightly more descriptive and acknowledged criteria are holistic rating schemes such as the ETS four or six-point scales ranking papers within a set. Teachers' use of a letter grade with some supporting comments might relate to this evaluation scheme. Some rating schemes are specific to discourse mode, others, like the primary trait rating method, are specific to discourse mode and the particular topic (Lloyd-Jones, 1977). The most detailed scales are analytic rating schemes referencing component features of the written product.

Where do these criteria come from? Criteria for these scales may be inferred from features commonly referenced by knowledgeable readers, may be arbitrary, or may be theoretically- or empirically-based dimensions deemed important by the group designing the scheme. Analytic scales vary in the degree to which they comprehensively reference rhetorical, structural syntactic features, as well as the degree to which criteria for features are qualitative, more objective, or, even, quantitative. In an attempt to be comprehensive, the subscales of the Diederich Expository Scale range from "ideas" to spelling (Diederich, 1974). In contrast, analytic text analysis schemes such as T-unit analyses or Halliday and Hasan's measures of cohesion focus on isolated components of the written piece (Halliday & Hasan, 1976). Diederich's "flavor" subscale is far more qualitative and judgmental than counts of numbers and types of cohesive ties. In classroom evaluations of student writing, grades and teachers' comments, too, may reference a range of essay features such as content, organization, and mechanics (Freedman, 1979); or their comments may only relate to sentence-level problems.

One issue in developing or using a rating scheme is the meaning of writing score(s). From a psychological perspective, does being a "2" vs. "4" discriminate between levels of a student's writing competence? At present, there is little research evidence that any sets of criteria used in actual practice are more valid than others for discriminating between levels of expertise. From a logical perspective, how specific, replicable, and informative are rating criteria? Pedagogically, what implications do the scores have for diagnosing strengths and weaknesses? The bases of the score, the criteria, should serve as feedback to teachers, students, and parents. To be fair, criteria employed in minimum competency tests should specify writing elements that are basic writing skills, e.g., organization, support, mechanics. The criteria should also be those amenable to instructional intervention. The more judgmental, qualitative, sophisticated and less teachable writing elements such as flavor, style, or voice would seem less fair, and useful, and would, therefore be inappropriate as rating criteria for judging basic writing competence. Specification of criteria may be the most important decision affecting the utility of information provided by assessment, both large scale and classroom level. Certainly, consensual decisions on these criteria should involve instructional and evaluation personnel.

It seems logical that criteria used in large scale writing competency assessment should reflect, if not derive from, criteria used to evaluate student classroom writing. An ideally integrated instructional system that targets particular writing elements as important basic competencies would involve teachers and evaluators in specification of rating criteria and encourage focused classroom guidance, feedback, and evaluation on these elements. Instructionally, specification of valued basic criteria could provide a more comprehensive framework for teachers to focus instruction and communicate feedback to students about their writing. The scanty research on classroom evaluation methods suggests that teacher comments more often cite easily identified sentence-level mechanical errors than text level feedback such as organization and support (Pitts, 1978; Quellmalz, Baker, & Enright, 1980). As Coffman pointed out, while few would recommend complete restriction and regulation of the criteria teachers use in classroom writing assessment, neither would they condone subjecting students and the instructional program to wildly fluctuating, idiosyncratic standards of individual teachers (Coffman, 1971). Some standardization of writing criteria seems particularly critical for minimum competency goals. And, of course, economically schools using criteria for systemwide assessment that are also used in classrooms would eventually, considerably reduce the cost of training raters.

Assuming that criteria have been specified that are logical, fair, and useful, the format for recording scores remains a problem. Many large scale assessments report a single, holistic score. A logical question is whether it makes sense to comment on component features of a student's writing instead of, or in addition to, its overall quality. A likely question to be raised about a single global score by a teacher, student, parent (or lawyer) is "Why?" followed by "Show me." While writing theory may suggest that the "whole" is greater than the sum of its parts, research in psychology and pedagogy suggests that learners advance when taught how to use components and combine them into competent performance (e.g., Skinner, 1957; Resnick, 1980). Another logical question is whether students are differently classified as masters and non-masters and/or if analytic schemes yield a differential score profile. Winters (1978) found that various scoring rubrics including a general impression scale, two analytic scales and a T-unit analysis, did classify students differently. Quellmalz, Smith, Winters & Baker (1980) found that three separate holistic rubrics and an analytic rubric classified entering freshman differently. Similarly, Polin (1980) found very low correlations between primary trait and analytic ratings of the same essays. Each of these studies compared scoring rubrics which referenced some similar criteria but which, in application, produced variable characterizations of the same essays. Still unexamined are the cost benefits of the scales using exact same criteria, but recording a single, holistic judgment or several separate analytic scores. Such a study is currently in progress (Quellmalz, 1981).

A major problem for large scale writing assessments, to be sure, is the cost of providing more detailed ratings. In the narrowest sense, cost is measured in terms of time required to train raters and time required to rate papers. Generally, training on more criteria that are more explicit requires more time than training on fewer or less explicit criteria.

Currently available data on scoring costs indicate that training time for holistic and primary trait scoring averages two to four hours (Powliss, Bowers, & Conlan, 1979; Mullis, 1980) and for analytic scoring averages six to eight hours (Smith, 1978; Quellmalz & Capell, 1979). Trained raters can assign a holistic or primary trait score to a student's paper reliably in 30 seconds to 1½ minutes (Powliss et al., 1979; Mullis, 1980). Rating time for providing five to eight separate analytic scores range from four to five minutes for multi-paragraph essays and from two to four minutes for paragraphs (Smith, 1978; Quellmalz & Capell, 1979).

In a recent study comparing two score formats, an analytic scheme or a holistic scheme modified to provide diagnostic checks for students rated below mastery, Quellmalz found that average rating times per paper differed by approximately one minute (Quellmalz, 1981). Is the additional training and rating time "worth it?" School systems weighing this question might consider broader definitions and implications of cost. First, the cost of either analytic or holistic training could be jointly shared as an in-service activity, by curriculum budgets. These training costs would also then decrease to review time when all teachers in a system were trained. A second potential cost sharing strategy is to view essay ratings as diagnostic components of the instructional system to both focus and monitor program improvement. A third cost concern is an ethical one. Students have spent considerable time producing writing samples and the psychological and opportunity costs to them of uninformative or erroneous classification as failures can be profound. Finally, a system might consider the degree of specific support useful for defending mastery/non-mastery classifications; the costs of remediation and lawsuits because of misclassifications can be high.

Problem 4: Technical Quality of Rating Criteria

A fundamental responsibility of an assessment program is the documentation of its technical quality. For writing assessments this becomes a problem of scale stability and validity, i.e., demonstrating that score criteria are applied uniformly within and between rating occasions and that other measures of student writing competence corroborate the test ratings (Quellmalz, 1980).

When carefully structured scale training sessions precede actual rating, most holistic and analytic rating scales can demonstrate high interrater reliability (Powliss et al., 1979; Mullis, 1980; Quellmalz, 1980; Steele, 1979; Van Nostrand, 1980). But interrater agreement within a rating session is not sufficient for demonstrating scale reliability. Analogous to the problem of test-retest reliability, a reliable scale must be stable, i.e., demonstrate that its criteria would be applied consistently by new sets of raters to both a new set of papers and to the set of papers scored by the first raters. To the extent that criteria are differently applied, the scale is not stable and reliable (Quellmalz, 1980).

Few scales currently used in writing assessment report data about their stability across sets of raters and rating occasions. It seems that scales with more explicit and operational criteria are less susceptible to fluctuating qualitative judgments and are more likely to be stable across paper sets and raters. Holistic scales such as the ETS method which awards scores according to a paper's ranking within a unique set of papers result in a sliding scale (Conlan, 1979). A "2" paper in one paper set may well have characteristics quite different from a "2" paper in a set of papers with a broader or narrower quality range. While some attempt is made to stabilize judgments across sets of raters by inserting anchor papers during training, anchor papers are less frequently interspersed in actual rating sequences. Statistical evidence of the comparability of scores given on any such anchor papers by different groups of raters is noticeably, and seriously, absent. Thus, holistic scales using ranking procedures within sets and unexplicated criteria are suitable for norm-referenced selection decisions, but can not meet competency test requirements for stable, uniform application of criteria. On the other hand, holistic scales based on more descriptive criteria such as the primary trait method (Lloyd-Jones, 1977) may be more likely to permit stable application across paper and rater sets. Reports for most analytic scales also document interrater reliability within rating occasion but do not track stability across occasions. For analytic as well as holistic scales, precision of criteria is a critical factor in achieving scale stability. School systems designing writing assessments should routinely report interrater reliability and check scale stability on common paper sets scored at different rating sessions. These measures will reassure stakeholders that assessments are uniform and fair.

The task of documenting the validity of writing assessment rating scales can take several forms. Most competency-based writing assessments attempt to establish content validity through expert judgments about the skills assessed (Breland & Ragosa, 1976). Few writing assessment programs subject the rating scales used to evaluate those skills to content validity scrutiny as well. Since, for written production, the scale defines what acceptable writing is, the content validity of scales should be judged by the same procedures as test items or specifications. It may be that some scales with vague criteria or criteria heavily weighted toward sentence-level mechanics would not get the stamp of approval from a broad range of experts. It should be noted that holistic scales with no explicit criteria are "content" free and assignment specific. These scales are not suitable for competency assessments.

Of course, content validity is only one index of validity (Cronbach, 1971; Messick, 1975). Both concurrent or predictive and construct validity should be examined. The most common method for validating large scale rating schemes has been to report their correlations with other writing-related measures including other English grades, reading test scores and multiple choice writing test scores. Many of these "criterion" variables, however, are even more questionable indicators of writing ability than the rating scale being validated. A major problem in validating rating scales is identifying appropriate criterion groups and test scores (Winters, 1978; Quellmalz, Spooner-Smith, Winters & Baker, 1980). A directly related criterion would be relationships of immediately preceding and subsequent writing assignment scores. Unfortunately, as different criteria are often employed in other rating scales and/or in teachers' grading of assignments, few appropriate direct comparisons are possible.

From the student's viewpoint, this problem raises concerns for fairness and instructional validity. How closely do the criteria used in the assessment match those used in the classroom, and how closely do they represent writing skills for which the student has received instruction? Fundamental precepts of fairness require that if a system hasn't explicitly taught the skills, it shouldn't hold the student accountable for being competent in these skills. For example, originality, humor, and flavor are desirable features of writing; they are not often taught directly. If we have no information on the criteria used in holistic scoring, that method isn't fair; we have no way to determine if what was tested was what was taught. The legal implications of this dilemma are obvious.

Summary

Balancing ideally detailed analyses of students' writing with the costs of those analyses is no easy task. School systems and teachers across the country are wrestling with the problem and arriving at varying solutions. Some systems don't even try to initiate large scale rating of writing samples. Some teachers assign little writing and provide cursory or global feedback. Other systems are willing to pay the price and mount articulated writing assessment and instructional systems (e.g., Detroit, Los Angeles, Pittsburgh).

Some rating schemes apply explicit, replicable, reasonable criteria; some scales are silly, some are misapplied, some are downright harmful.

Large scale assessments can devise ways to reduce the costs of training raters to score large numbers of essays. In an ideally integrated assessment system, tasks and criteria for the large scale assessment would be the same as those used in the classroom. A district or state might construct a scale that referenced basic text components used by classroom teachers, e.g., main idea, coherence, support, mechanics, and devise a scoring system where papers were checked off as competent on each skill and also check off in more detail the components falling below mastery. For example, one text might have competent support and receive a mastery check; another essay might not and get a check for "details are not related to the main point," or "details are not concrete."

Systems might allocate the cost of training raters to staff development. All system teachers could be trained in applying the rating criteria which should promote greater articulation of the formal assessment with classroom criteria. Districts such as Detroit find it cost effective to pay lay personnel to rate writing samples. Alternately, the system might ask teachers to swap papers. Teachers could use the rating scale to score writing of other students in the district in return for having their students' writing scored by other teachers trained as raters. This would reduce training costs for district scoring. Many alternative logistics could be engineered to spread the time and energy costs efficiently within existing system resources.

References

Baker, E. L., & Quellmalz, E. S. Issues in eliciting writing performance: Problems in alternative prompting strategies. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA, April 1980.

Bourne, L. J. *Human conceptual behavior*. Boston: Allyn & Bacon, 1966.

Breland, H. M., & Braucher, J. L. Measuring writing ability. Paper presented at the annual meeting of the American Educational Research Association, New York, 1977.

Breland, H., & Ragosa, D. Validating placement tests. Paper presented at the annual meeting of the American Educational Research Association San Francisco, 1976.

Britton, J. The composing process and the functions of writing. (Chapter 2) In Cooper and Odell (Eds.), *Research on Composing: Points of departure*. Urbana, IL: National Council of Teachers of English, 1978.

Cazden, C. B. Two paradoxes in the acquisition of language structure and functions. In K. Connolly & J. S. Bruner (Eds.), *The Growth of Competence*. New York: Academic Press, 1974.

Coffman, W. E. Essay exams. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Conlan, G. *Comparison of analytic and holistic scoring techniques*. Princeton, NJ: Educational Testing Service, 1979.

Cooper, C. R. Current studies of writing achievement and writing competence. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.), Washington, D. C.: American Council on Education, 1971.

Crowhurst, M. Syntactic complexity in narration and argument at three grade levels. *Canadian Journal of Education*, 1980.

Diederich, P. B. *Measuring growth in English*. Urbana, IL: National Council of Teachers of English, 1974.

Florio, S. Learning to write in the classroom community: A case study. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Follman, J. C., & Anderson, J. A. An investigation of the reliability of five procedures for grading English themes. *Research in Teaching of English,* 1967, 190-200.

Freedman, S. How characteristics of student essays influence teachers' evaluation. *Journal of Educational Psychology*, 1979.

Godshalk, F. I., Swineford, F., & Coffman, W. E. *The measurement of writing ability*. New York: College Entrance Examination Board, 1966.

Halladay, M. A., & Hassan, R. *Chohesion in English*. London: Longman, 1976.

Lloyd-Jones, R. Primary trait scoring. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging*. Urbana, IL: National Council of Teachers of English, 1977.

Messick, S. A. The standard problem: Meanings and values in measurement and evaluation. *American Psychologist*, 1975, pp. 955-966.

Moss, P., Cole, N., & Khampalikit, C. A comparison of direct and indirect writing assessment methods. *Journal of Educational Measurement*, in press.

Mullis, J. A. Using the primary trait system for evaluating writing. National Assessment of Education Progress, 1979.

National Assessment of Educational Progress. Reading, thinking and writing: Results from the 1979-80 National Assessment of Reading and Literature. Denver, Colorado, 1981.

Pitts, M. Relationship of classroom instructional characteristics and writing in the descriptive/narrative mode. Center for the Study of Evaluation, University of California, Los Angeles, CA, November, 1978.

Polin, L. Alternative conceptions of the writing skill domain: Problems for the practitioners. Paper presented at the National Council on Measurement in Education, Boston, 1980.

Powills, J. A., & Bowers, R., & Conlan, G. Holistic essay scoring: An application of the model for the evaluation of writing ability and the meaurement of growth in writing ability over time. Paper presented at the annual meeting of the American Educational Research ASsociation, San Francisco, 1979.

Praeter, D., & Padia, W. Effects of modes of discourse in writing performance in grades four and six. Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

Quellmalz, E. S. Domain-referenced specifications for writing proficiency. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1978.

Quellmalz, E. S. Assessing writing proficiency: Designing integrated multi-level information systems. Paper presented at the annual meeting of the National Reading Conference, San Diego, CA, 1980.

Quellmalz, E. S. Report on Conejo Valley's Fourth-Grade Writing Assessment: Fall, 1981.

Quellmalz, E. S., & Capell, F. Defining writing domains: Effects of discourse and response mode. Report to the National Institute of Education, November, 1979. (Grant No. OB-NIE-G-78-0213 to the Center for the Study of Evaluation.)

Quellmalz, E., Spooner-Smith, L. S., Winters, L., & Baker. Characterizations of student writing competence: An investigation of alternative scoring systems. Paper presented at NCME, April 1980. (Grant No. OB-NIE-G-79-0213 to the UCLA Center for the Study of Evaluation, 1980.)

Quellmalz, E., Baker, E., & Enright, G. Studies in Test Design: A Comparison of Modalities of Writing Prompts. Center for the Study of Evaluation, University of California, Los Angeles, CA, November, 1980.

Resnick, L. What do we mean by meaningful learning? Invited address at the annual meeting of the American Educational Research Association, Boston, 1980.

Skinner, B. F. Verbal Behavior. New York: Appleton, 1957.

Scribner, S., & Cole, M. Unpackaging literacy. Social Science Information, 1978, 17, 19-40.

Smith, L. S. Investigation of writing assessment strategies. Report to the National Institute of Education. Center for the Study of Evaluation University of California, Los Angeles, November 1978.

Steele, J. M. The assessment of writing proficiency via qualitative ratings of writing samples. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Van Nostrand, A. D. Writing Instruction in the Elementary Grades: Deriving a model by collaborative research. Providence, RI: Center for the Research in Writing, 1980.

Veal, L. R., & Tillman, M. Mode of discourse variation in the evaluation of children's writing. Research in the Teaching of English, 1971, 5, 37-45.

Winters, L. The effects of differing response criteria on the assessment of writing competence. Report to the National Institute of Education, November 1978. (Grant No. OB-NIE-G-78-0213 to the Center for the Study of Evaluation.)

THE MEASUREMENT OF STUDENTS' WRITING PERFORMANCE IN RELATION TO INSTRUCTIONAL HISTORY

Marcella Pitts

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

X2CSE/B

THE RELATIONSHIP OF CLASSROOM INSTRUCTIONAL CHARACTERISTICS AND WRITING IN THE DESCRIPTIVE/NARRATIVE MODE

Public concern over an apparent decline in students' writing skills has prompted educators to examine two central issues: (1) the design of composition curricula and (2) the valid and reliable assessment of students' writing performance. This study addressed these two issues by describing instructional characteristics in a specific curriculum, by developing and employing an analytic rating scale to evaluate students' writing performance, and by examining the relationship of instructional characteristics to writing performance.

This research was exploratory in nature and exhibits limitations inherent in exploratory studies such as small sample size. Nevertheless, the study provided descriptive information about selected instructional characteristics of composition classrooms and, thereby, provided data relevant to current concerns about composition curricula. Secondly, generalizable procedures for constructing and field testing an analytic rating scale and for training raters in its use were obtained from this research and may contribute to our knowledge base in the area of writing assessment.

Among recent developments in our efforts to calm public anxiety over students' writing skills and to discern how best to teach writing and assess students' writing are the design and implementation of composition courses incorporated as requirements into the high school curriculum. This research focused on a typical curriculum change designed to improve writing skill, a one-semester required composition course developed by a large urban school district and incorporated into the curriculum at the eleventh grade.

Briefly, the course provides instruction in four domains of writing: (1) the sensory/descriptive, (2) the imaginative/narrative, (3) the practical/informative, and (4) the analytic/expository. The recommended minimum number of compositions for each of the domains is three, making the minimal number of completed compositions for the semester 12. Teachers are encouraged to offer instruction in each of the domains and to include in their instruction: (1) prewriting and precomposing activities to elicit ideas from students and to motivate them to write; (2) writing practice to increase flexibility, fluency, skill, and confidence; (3) reinforcement; and (4) instruction in grammar as it relates to the writing process.

The course's curriculum outline and these recommended activities were valuable resources in the design of two instructional questionnaires, the primary data collection instruments in the study.

Information concerning instructional practices was obtained from teachers and students for a selected group of variables: (1) communication of instructional outcomes to students, (2) writing practice, (3) feedback, (4) instructional time use, and (5) teacher expectation. In addition, papers previously assigned and graded by the teachers supplied information about the usual emphases and specificity of correction provided students.

Foremost in the selection of these variables over other instructionally important dimensions identified in the literature was the fact that they involve concrete instructional events. The presence, absence, and frequency of occurrence of these events can be monitored and reported by teachers and students. This was an important consideration given the methodology used in the study, which relied heavily on teacher and student self-report.

Students' writing performance was measured by their combined scores on two narrative/descriptive writing tasks. An analytic rating scale, developed for the study and appropriate for the narrative/descriptive mode, was employed by three high school teachers to rate the writing samples. The teachers, all of whom had rated essays previously, were trained in the use of the rating scale.

Sample

The subjects of the study were the students and teachers in 19 composition classrooms in five high schools in a large urban school District. The selection of the schools in the sample was based on achievement and demographic data published annually by the district. These data were used to develop profiles of individual high schools in the district; the five schools selected for inclusion in the study had relatively homogeneous profiles along these dimensions.

The number of classes participating ranged between three and four per school. Participation was voluntary, with the decision to take part in the study resting with the individual teachers. Six of the classrooms were designated by the participating schools as advanced (above average); 11 as average; and two as skill (below average) classes.

Procedure

Data collection in the five schools took place during the last two weeks of May 1978. Visits to each school were scheduled to provide for an interval of approximately one week between writing assignments. Forty minutes writing time was allotted for each writing occasion; order of topics was counterbalanced by class.

During this period teachers provided the investigator with a set of previously graded student compositions. After the writing samples and sets of graded papers had been collected, students and teachers completed the instructional questionnaires.

All the essays were returned to the students at the completion of the study. Several teachers used the essay as a graded class assignment.

Independent Variables

The independent variables in the study were: (1) communication of instructional outcomes, (2) use of instructional time, (3) writing practice, (4) feedback, and (5) teacher expectation. Information related to each of these variables was collected from teachers and students via questionnaires. Parallel items pertaining to many of the variables appeared on both the teacher and student questionnaires.

The first independent variable, communication of instructional outcomes or intent, was operationally defined as informing students of the skills they were expected to acquire at the end of the semester. In order to ascertain the extent to which teachers had successfully communicated this information, teachers and students were provided with parallel lists of post-instructional skills and asked to select those which matched most closely the instructional outcomes in their classrooms. An index which measured the agreement between the skills selected by students and teachers was then computed.

The second independent variable, time on academic content, was operationally defined as the amount of time spent on: (1) modes of discourse, (2) writing activities, and (3) features of writing measured by the analytic rating scale developed for the study. Data for the first and second dimensions of this variable were purely descriptive. They involved teachers' estimates of the percentage of instructional time spent on each of the four domains of writing in the curriculum and teachers' and students' reports of the activities on which class time was spent (e.g., reading composition texts, reading literature, prewriting discussions, in-class writing).

The third dimension of this variable was measured by questionnaire items which required teachers and students to indicate the number of class periods (i.e., 0, 1-5, 6-10, over 10) spent on specific features of writing related to description, narrative order, and mechanics.

The third independent variable, practice, was defined as the amount of writing, i.e., frequency and length of assignments written in class and as homework. Another dimension of the practice variable for which descriptive data were obtained was the type and number of writing assignments students completed over the semester. Teachers and students estimated the number of completed compositions that were expository, descriptive, narrative, multi-paragraph, single-paragraph, etc.

The fourth variable, feedback, was composed of three dimensions: immediacy, helpfulness, and instructiveness. The first dimension, immediacy of feedback, was measured by parallel questionnaire items in which teachers and students estimated the time period within which students' papers were usually read and returned. Since the results of previous research and data from pilot tests had indicated that the most common methods of providing feedback were individual conferences with students and written comments on compositions, information was obtained from students concerning these evaluation practices. Helpfulness of written feedback on corrected papers and feedback received during individual conferences with the teacher were measured by students' ratings. An item on the student questionnaire also provided a measure of the instructiveness of feedback, and additional data related to this dimension were obtained in a separate analysis of sets of previously assigned and corrected papers provided by the teachers. In addition, teachers reported on the features on which they focused during correction and the usual methods they used to provide feedback.

The fifth independent variable in the study, teacher expectation, was measured by teachers' recollections of the amount of improvement they had expected in students' performance at the beginning of the semester.

Dependent Measure

The writing task in the study was closely related to the sensory/descriptive and imaginative/narrative domains described in the curriculum outline. The task, primarily but not purely narrative, was structured so that descriptive detail would be included in the two compositions students wrote.

Both of the writing assignments used pictures as writing stimuli. Students were directed to write about the scene in the pictures and the events which might have preceded and followed it. They were to include in their essays descriptive detail for readers who would not have an opportunity to look at the pictures. Further, students were directed to use the third person point-of-view. They were told that the purpose of the assignment was to write a story based on the picture and to include descriptive detail related to setting, characters, and action.

The selection of the pictures and the development of the directions for the two assignments were based on data from field tests with comparable students.

A narrative/descriptive analytic rating scale was developed and used to evaluate students' essays. The features included on the scale were derived from a survey of theoretical and practical works on descriptive and narrative writing as modes of discourse and from an examination of published rating scales. Draft versions of the scale were reviewed by faculty in the UCLA English department and by staff at the Center for the Study of Evaluation. The final version of the scale reflected the changes suggested by the reviewers as well as minor modifications agreed upon by the investigator and the readers prior to the actual rating of students' essays.

As a first step in developing the scale, a review of available theoretical and practical pieces on descriptive and narrative writing was conducted. This review resulted in the identification of four essential features of narrative/descriptive writing which appeared to be appropriate criteria for evaluating relatively short pieces written under timed conditions: setting, characterization, action, and descriptive detail, the inclusion of which contributes to the reader's sense of setting, the characters, and the action or sequence of events.

The selection of these features was supported by a state-of-the-art review of published analytic rating scales relevant to narrative/descriptive writing. A review of non-mode-specific features of these and other prominent analytic rating scales was also undertaken to identify elements related to mechanics for inclusion in the scale.

Based on this review, a sentence-structure/diction subscale and a grammer/spelling subscale were developed. The criteria on the first scale include fluency and variety of sentence structure, the selection of clear and specific words and their correct use. The grammar dimension of the grammar/spelling subscale focuses on reference errors, tense shifts, punctuation errors, misplaced modifiers, and the like. The first mode-specific subscale, sequence/coherence, includes criteria related to temporal order of events, their logical development, and the continuity with which they are developed. While the criteria for this subscale focus on the narrative aspects of the writing task, the criteria for the other mode-specific features, setting and characterization, focused more on the descriptive aspects of the task.

Students' essays (n = 228) were rated by three trained readers for each of the features on the narrative/descriptive scale. All the readers were high school English teachers; all had prior experience in holistic rating. Approximately two days were spent training the readers and another one and one-half days were required to read and rate the essays.

Prior to the training and rating sessions, all identifying information was removed from the students' essays and each paper was assigned a code number. All the essays were then typed to facilitate rapid reading and to remove the confounding effects of handwriting. No corrections of any kind were made on the typewritten versions: They were duplicates of the handwritten essays in every respect. In addition to the 228 sample essays, 140 extra essays were prepared in a similar manner for use in rater training and in calculating inter-rater reliability.

Both days of rater training were full-day sessions. The morning of the first day was spent reading, discussing, and applying the rating scale to specially selected training essays chosen to represent the range of essays in the sample. A procedure was followed in which the readers rated one, two, or three essays individually and then discussed their ratings. During the discussions discrepant ratings were examined; elements on the scale and the terms used to describe them were clarified. A first inter-rater reliability check was conducted during the afternoon. A three-way analysis of variance (ANOVA) was used to calculate inter-rater reliability for the ratings assigned to 40 essays (two essays written by each of 20 students) read over a one and one-half hour period. The three factors in the analysis were subjects, topics, and raters. A mixed model was employed in which subjects and raters were random factors and topics were fixed.

The second day of training was spent discussing, refining, and applying the subscales for which the initial reliability coefficients had been considerably below the .80 level for average ratings chosen as the test of acceptability. The rating of the 228 essays began when the inter-rater reliability coefficients for each subscale had increased and were greater than .80.

The essays were placed in a different random order for each rater so that the likelihood of all readers rating the same essay at the same point in time was reduced. However, a common group of essays was included in each reader's stack of papers so that another reliability check could be conducted.

The final inter-rater reliability for the 228 essays was quite high, with average ratings of .88 and .89 and single ratings of .72 and .74. The total score reliabilities were .94 (average ratings) and .87 (single ratings).

On the average, student performance was not high or low on any one subscale. Furthermore, the mean performance of students in classes designated as above average was consistently higher than the mean performance of students in average and below average classes. The mean performance of average level students was, in turn, consistently higher than the mean performance of students in below average classes. In all cases writing performance was measured by total writing score for each student or mean total writing score in each classroom.

Data Analysis and Results

A two-stage data analysis was performed to examine the relationships between students' and teachers' reports on classroom use of instructional variables and the quality of student writing samples.

In the first stage of the analysis, descriptive statistics were performed on the data from the teacher and student questionnaires. The research questions addressed were:

1. How are selected instructional variables employed in the composition classroom: (a) as reported by students? (b) as reported by teachers?
   - Are intended instructional outcomes communicated to students?
   - How do teachers allocate instructional time to writing activities, modes of discourse, and specific writing skills?
   - What kind and how much opportunity for writing practice is provided?
   - What is the time interval for feedback, what form does it take, and how instructive is the feedback provided students?
2. What expectations do teachers have concerning students' writing performance?

In the second stage of the analysis, a series of multiple regressions was performed to examine the relationships among reported use of the independent variables and student writing performance. The research questions addressed were:

1. What is the relationship between students' writing performance and use of the four instructional variables: (a) as reported by students? (b) as reported by teachers?
2. What is the relationship between students' writing performance, use of the four instructional variables, and teachers' expectations?

Results of the descriptive analysis of the questionnaire data provide a rich base of information concerning instructional practices in the 19 classes in the sample.

According to the teachers and students in these classes, the intended instructional outcomes in the majority of the classrooms were: to write complete and grammatically correct sentences; to write well-organized essays; to include supporting detail in essays; to use a consistent point of view in writing; to follow accepted standards of usage; and to express ideas in an original way. Students in above average level classes were most in concert with their teachers regarding the instructional outcomes in their classrooms.

With respect to the classroom activities designed to achieve these outcomes, teachers and students in the majority of the classrooms agreed that activities were prewriting discussion, in-class writing, composition analysis, reading literature, and listening to formal lectures by the teacher. The first three activities are recommended in the course's curriculum guide.

The guide also suggests that teachers spend approximately equal time during the semester on each of the four writing domains specified in the curriculum. In fact, eight of the teachers in the sample indicated they divided their available instructional time equally among the four domains. The majority of these teachers taught above average classes.

With respect to the type of assignments completed, the majority of the teachers and their students agreed that teachers offered one-to-five assignments for narrative, descriptive, expository, and argumentative writing over the semester. Also, more than half of the students indicated they had written one-to-five short stories during the course of the semester. Teachers of average level classes assigned more grammar exercises than their counterparts in above average classes, while this latter group assigned more research papers and multi-paragraph essays.

As might be expected, all the students in the sample wrote in class more frequently than at home. Teachers of above average classes, however, had their students engage in writing activities more often, both at home and in class, than did average level teachers. Moreover, the in-class essays of above average students were longer than those required of average level students.

Above average class teachers also spent more time in individual conferences with students and provided more specific rules and suggestions for improvement in the written comments on students' papers. Less than half of the teachers of average classes had individual conferences with their students to discuss an assignment. They also wrote fewer directive comments on students' papers and, as might be expected, had a faster turnaround time for corrected papers.

In an analysis of the comments on previously graded papers provided by the teachers, the comments on over one quarter of the sets were rated as highly directive since they included specific rules and suggestions for students. The majority of these papers were from above average level classes. Slightly less than one quarter of the sets provided specific indications of strengths and weakness but failed to suggest specific strategies to improve the paper. A similar percentage contained no comments; and the comments on the remaining papers, nearly one quarter of the total, were too general to be of any instructive value and contained only general remarks about the paper.

In addition, teachers reported they attended to content and mechanics or organization and mechanics when they corrected students' papers. As might be expected, the analysis of both the interlinear notations and comments on the sets of previously graded papers showed that more notations pertained to mechanics.

The results of the descriptive analysis indicate that important differences may exist in instruction between competency levels. The pattern of instruction in above average level classes seemed to rely upon and extend students' initial writing skills. Students wrote more often, wrote longer essays, had more individual conferences with their teachers, and received more instructive feedback than did students in average level classes. Not only did teachers of average level classes make shorter assignments, these included more grammar practice. Grammar exercises are de-emphasized in above average classes. Thus, the data suggest that these teachers teach to the competency level of their classes. More competent classes receive more demanding instruction; less is expected and asked of less competent groups. Indeed, teachers' expectations concerning the amount of improvement in students' writing performance over the semester is positively and significantly correlated with the school-designated competency or tracking level of the classes.

The powerful influence of tracking level on student performance was more apparent in the regression analyses performed. Results of the multiple regressions based on students' reports, teachers' reports, and the discrepency between the perceptions of both groups indicated that, for the variables examined in this study, classroom tracking level is the single significant variable related to students' writing performance. Thus the analyses revealed no relationship between reported instructional practices and students' performance, despite the findings that such practices tend to vary for classes in different tracking levels.

These results may derive from limitations in the design and scope of the study and from additional constraints imposed by the curriculum itself. Nevertheless, they inform further research and invite secondary analysis.

Because of the exploratory nature of this study, the sample size, using the classroom as the unit of analysis, was extremely small (n = 19). When the sample was further subdivided into tracking levels, numbers were even smaller: average classes, n = 11; above average classes, n = 6; below average classes, n = 2. Within the limitations of the sample size, it was impossible to examine the relationship between total writing score or subscale scores and instruction due to the correlation between instruction and tracking level.

An additional constraint which may have hampered the discovery of significant relationships lies in the curriculum itself. This course is only a semester long, and yet the curriculum requires instruction in each of four writing modes. As reported in the findings above, teachers do in fact provide instruction in all four modes. Furthermore, there is some indication that the course is more of a survey of different writing domains than extensive drill. For example, the curriculum recommends that a minimum of three assignments be completed in each of the four domains over the semester. Given that the mission of this course is to provide students with basic writing competencies if they have not mastered these in previous classes, this number appears to be quite conservative. The results showed that teachers provided a moderate amount of writing practice and moderate number of writing assignments in contrast to the more intensive instruction that might be expected in a composition course of this nature. Perhaps limiting the curriculum to fewer modes of discourse or expanding the course length to one full year to accommodate all four modes would strengthen instructional effects. Future studies under such conditions might uncover relationships that were too weak to appear in the present study.

Despite these limitations, other methodological features of the study appear to be promising strategies for studies of this type. First, the collaborative reports of teachers and students provided a reasonable indicator of instructional practices. Teachers and students, especially those in above average classes, were in considerable agreement, and the collection of survey data from both of these groups seems feasible and practical, especially at the senior high level. Future work might make use of more frequent surveys throughout the semester to prevent honest inaccuracies in recalling information over a long period of time. Studies should also include direct observation of classrooms. Observation of ongoing classroom interactions and instructional processes would allow more precise description of instruction and corroboration of questionnaire data.

A second promising strategy included in this study which could be incoporated into future work was the examination of teachers' naturally occurring comments as a way to qualify their self-report data. Other procedures of qualifying the data provided in this type of study should be examined as well.

Another product of the study which can be applied in other research studies is the narrative/descriptive analytic rating scale developed for the study. Experienced readers with a minimal amount of training can achieve highly reliable ratings using this scale. Moreover, it appears to be a valid measure of writing performance given the high correlation between tracking level and the mean total writing score of students in a particular classroom.

APPENDIX

NARRATIVE/DESCRIPTIVE RATING SCALE

1. Sequence/Coherence--criteria for rating:

1 point	There is no clear temporal (chronological) order to the events in the narrative. The reader is not sure which event comes first or follows any other event. In fact a sequence may not be related at all. The paper may be purely expository or descriptive.
2 points	There is a noticeable beginning and end although the temporal order of events may not be clear. Events are merely listed rather than progressively and logically related to each other. Sentences and paragraphs are poorly tied together. There are lapses in coherence; or, if transitions are used, they may be used incorrectly or repetitiously.
3 points	The temporal order of events is clear. Transitions are used correctly. The paper has continuity and there is a clear progression of ideas, although there may be minor lapses in motivation and logic.
4 points	This paper has all the elements of a "3" paper, with the addition of a sense of control from beginning to end. The sequencing of events is so well done that the reader has a sense of movement. There is a logical progression of ideas. Transitions may be expertly used and movement facilitated by a variety of transitions. The paper is often interesting, original and may include conflict.

II. Setting--criteria for rating:

1 point	The setting of the narrative is not clear to the reader because: (1) the writer does not specify where or when the action is taking place; (2) the reader is unable to infer the setting from the information included in the narrative. The setting is so vague, general and unspecific that the reader has no image of time or place.
2 points	The setting of the narrative is apparent to the reader. The actual place or time is stated or inferred, but there is little or no elaboration.
3 points	The reader has a clear understanding of setting. The setting is more explicitly stated than in a "2" paper. Details may relate to geographic location, time period, or general environment through which the characters move.
4 points	This paper includes all the elements of a "3" paper, with the addition of excellent use of detail. The writer uses specific detail to describe the setting. The setting is so developed that it seems to give the events a "real" place in which to happen. The setting is an important component in furthering the narrative.

III. Characterization--criteria for rating:

1 point	Characters are not identified or only barely identified: by name, noun, pronoun or there may be one or more adjectives which act like labels. However, there is no conscious attempt to develop the characters through their speech, actions, reactions to other characters or other characters' reactions to them.

2 points	Compared to a "1" paper, this narrative includes more information about one or more characters but this information is not elaborated. Details may only be listed, not developed. Characters are not clearly established.
3 points	Detail, interpretative comments, specific actions and reactions of the characters may be included. One or more of the characters may be a stereotype. Character is established and a specific direction for development is indicated.
4 points	One or more of the characters in the narrative may emerge as a unique, attention-getting person. A specific character is well-developed through dialogue, action, reactions to other characters, or by descriptions and/or interpretations of the character's appearence, feelings, or thoughts.

IV. Sentence Structure/Diction--criteria for rating:

1 point	Sentences are garbled, incomplete. Numerous structural problems interfere with the reader's comprehension. The sentences are not coherent; words are merely strung together. Monosyllabic words are used and the vocabulary is childish.
2 points	Sentences may be short and choppy or run-on. There may be fragments and comma splices. Word choice is limited. Words may be used incorrectly, repetitiously and inaccurately.
3 points	The sentences read without noticeable breaks and there is variety in sentence structure. There may be some sentence

errors but the paper is fluent. Word choice is exact and appropriate although uninspired. There may be several cliches and overworked expressions. The paper may be stilted or inflated.

4 points The paper has mature sentences making it easy and pleasing to read. It is marked by strong and precise diction. Vivid descriptive words which suit the writer's purpose are used.

V. Grammar/Spelling--criteria for rating:

1 point There are numerous grammatical errors (e.g., agreement, pronoun reference, misplaced modifiers, tense shift, punctuation) which interfere with the paper's readability. The writer seems to have no grasp of basic spelling rules.

2 points This paper is readable although the grammatical errors are distracting. There are several spelling errors in common words.

3 points The paper is basically competent. Errors are noticeable but they do not interfere with the writer's message. Spelling errors occur in words that are harder to spell.

4 points This paper has very few or no grammatical or spelling errors. The errors that remain make little difference to the reader; they are editorial problems and slips.

References

Braddock, R., Lloyd-Jones, R., & Schoer, I. Research in written composition. Urbana, Illinois: National Council of Teachers of English, 1963.

Conlan, G. How the essay in the CEEB English test is scored. Princeton, N.J.: Educational Testing Service, 1976.

Follman, J. C., & Anderson, J. A. An investigation of the reliability of five procedures for grading English themes. Research in the Teaching of English, 1967, 190-200.

Hambleton, R., Swaminathan, H., Algina, J., & Coulson, D. Criterion referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1977, 48(1).

Harris, C. W. Some technical characteristics of mastery tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced testing. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Howerton, M. C., Jacobson, M. & Eldon, R. The relationship between quantitative and qualitative measures of writing skill. Paper presented at the Annual Meeting of the American Educational Research Association, New York, April 1977.

Hunt, K. W. Early blooming and late blooming syntactic structures. In C. Cooper & L. Odell (Eds.), Evaluating writing. State University of New York at Buffalo, 1977.

Lloyd-Jones, R. Primary trait scoring. In C. Cooper & L. Odell (Eds.), Evaluating writing: Describing, measuring, judging. Urbana, Ill.: National Council of Teachers of English, 1977.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, Calif.: McCutchan, 1974.

O'Hare, F. Sentence combining: Improving student writing without formal grammar instruction. NCTE Research Report No. 15. Urbana, Ill.: National Council of Teachers of English, 1973.

MEASURES OF HIGH SCHOOL STUDENTS' EXPOSITORY WRITING: DIRECT AND INDIRECT STRATEGIES

Laura Spooner Smith

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California 90024

The project presented or reported herein was performed pursuant to a grnat from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

X2CSE/C

ABSTRACT

As demand increases for competency-based tests of students' basic adacemic skills, additional requirements for measures of writing proficiency also have surfaced. The need today is for measures of writing that not only are technically sound, but which also serve as meaningful, efficient indicators of clearly defined writing competencies. Additionally, the demand is for measures that carry clear implications for instructional planning.

MEASURES OF HIGH SCHOOL STUDENTS' EXPOSITORY WRITING: DIRECT AND INDIRECT STRATEGIES

As demand increases for competency-based tests of students' basic academic skills, additional requirements for measures of writing proficiency also have surfaced. The need today is for measures of writing that not only are technically sound, but which also serve as meaningful, efficient indicators of clearly defined writing competencies. Additionally, the demand is for measures that carry clear implications for instructional planning.

The present study was undertaken in an effort to examine relationships among writing assessment strategies which are potentially responsive to requirements of competency-based testing. Three alternate strategies to measure secondary students' expository writing were developed and administered. Two of the strategies, direct measures, involved collecting and rating students' writing samples. The distinction between the direct measure strategies lay in the response criteria by which the samples were judged. One form of criteria, an Analytic rating scale, required raters to assign scores to six different characteristics of the writing samples. The other form of criteria, an Impressionistic rating scale, yielded a single score on the quality of each essay as an example of exposition. The third strategy, an indirect measure, was an objective test of writing-related competencies derived from the Analytic rating scale.

Background

The measurement of writing and writing-related competencies historically has presented unique technical and validation problems. From pioneer assessment efforts to current composition tests, measurement experts have contended with such recurring problems as fluctuating test reliability, implementation of efficient scoring procedures, and examining the validity of indirect, i.e., objective, measures of writing.

In recent years, additional requirements for measures of writing have arisen along with increasing demand for competency-based tests of basic skills. This situation has prompted renewed attention to a fundamental concern in writing assessment: How to identify and define measurement tasks that will serve as efficient and instructionally meaningful indicators of writing competence. Identifying test tasks (or items) to measure a domain of learning is, of course, an intellectually difficult aspect of competency-based assessment, especially for complex behaviors such as the production of written discourse. Millman's (1974) assertion that a performance domain should be defined ". . . by those facets and elements that make a difference in how the learner responds" leads naturally to the question, what facets make a difference?

The Response Criteria Issue

In the measurement of writing skills, one "facet" that clearly makes a difference is the criteria by which writing samples are judged. Although a variety of guided scoring procedures for sorting or ranking written pieces currently are in use, most can be classified as either "analytic" or "holistic."

Analytic scoring procedures presume that a piece of writing can be viewed as consisting of component, but not necessarily independent, parts which are worthy of individual scrutiny. The procedures require raters to assign points to each of several specified aspects of a composition and yield an estimate of overall quality of the writing product, as well as sub-scores on separate elements of the writing sample. Analytic rating procedures aptly meet the requirements of competency-based assessment programs which call for information on specific skill strengths and deficits.

In contrast, holistic rating procedures assume that ". . . each factor that makes up writing skill is related to all other factors and that one factor cannot be easily separated from others" (Office of the Los Angeles County Superintendent of Schools, 1977). Under holistic procedures, raters assign a single value to a piece of writing. A specific type of holistic procedure, impressionistic rating scales, also requires assignment of a single value to a written product. Unlike most holistic approaches, impressionistic scales involve only a minimal rubric to guide the judgments of raters. The apparent advantage of holistic rating procedures in general, and impressionistic scales in particular, is that they tend to be an efficient and reliable direct measure of writing performance. The drawback is the lack of precise information on the attributes of writing to which raters attend. Whether analytic and holistic procedures provide substantially comparable estimates of writing competence is an important question, especially within the context of competency-based assessment.

The Response Mode Issue

Another "facet" in the measurement of writing skills is the form of response elicited from examinees. Indeed, an important issue in writing assessment centers on the relative merits of two contending response modes: production of a piece of writing, i.e., "writing samples," and selecting a response from among given alternatives, i.e., "objective" tests of writing.

That there is any question regarding the most appropriate approach for assessing composition skills at first may appear puzzling. Clearly, writing samples represent a direct measure of writing performance and therefore possess _prima facie_ validity. This, in fact, is the principal argument supplied by proponents of writing samples as a means of composition assessment. Those who favor more indirect methods, i.e., objective tests, however, have pointed to the usual unreliability of writing samples, notably the difficulty of ensuring reliable scoring procedures and the tendency of a writer's performance to fluctuate in quality over time and task. Problems of unreliability in writing samples have been the subject of recent research which has resulted in promising procedures aimed at enhancing measurement reliability. Nonetheless, a clear advantage of objective methods of assessment is the efficiency with which the measures can be administered and scored, an economy of especial importance when sizable aggregates of students are to be tested.

Arguments of efficiency, though, have failed to impress many professionals in the discipline of English, who have voiced concern over the inherent lack of content validity of indirect approaches. In response to such objections, proponents of indirect measures have cited findings from studies (e.g., Godshalk et al., 1966) which reveal statistically significant correlations between performance on objective tests of writing and performance on actual writing samples.

The limitation of many correlational and predictive studies of objective tests of writing for competency-based assessment, however, centers on the criterion against which the items are validated, notably the use of holistic scoring procedures to rate the criterion essays. Recall that holistic scoring results in a single score on the overall quality of an essay: separate scores on different characteristics of an essay are not provided. Consequently, no inferences can be drawn about possible relationships between the classes of skill tapped by objectve items and the classes of skill exhibited in the criterion essay. At best, one can conclude that various combinations of skill measured by objective items are related in some unspecified fashion to actual writing performance. What appears to be needed is information on the relationships between well-defined classes of objective items and equally well-defined writing production measures.

Method

Subjects, 128 eleventh and twelfth grade students in six English classes in the Los Angeles area, were randomly assigned within each class to treatment groups determined by the order in which the measures were administered. Table 1 depicts the three measurement strategies. Each subject wrote two essays of at least 200 words on topics designed to elicit expository writing and completed an objective test of writing-related competencies. Two raters were trained to employ an Analytic rating scale and two to use an Impressionistic rating scale. The writing samples were scored by both rater pairs, resulting in four total scores for each sample. Final study reliability of ratings on the Analytic total scale was .90, with rater reliabilities for the six Analytic subscales ranging from .84 to .95. Rater reliability for the Impressionistic scale was .87.

Measures

The direct measure writing task employed in the study consisted of two major components: the writing topics (and directions) and two forms of rating criteria. The topics, designed to elicit like-samples of student writing, were intended to promote writing within the discourse domain of exposition, that is writing ". . . that explains or clarifies a subject" (Brooks & Warren, 1961). Task attributes guiding development of the topics included discourse mode, rhetorical purpose, content limits, and intended audience.

One form of rating criteria, the Analytic scale, was designed to reflect state-of-the-art pedagogical precepts and practice in composition. Elements on the scale were derived from an analysis of conventionally recognized structure features of exposition as indicated in curriculum guides and textbooks at the secondary level. The final version of the Analytic scale yielded scores on six subscales corresponding to the following elements of writing: essay focus (main idea), organization, development, support, paragraphing, mechanics. The range of points for each of the Analytic subscales was four (high) to one (low). The rating rubric contained descriptions of essay characteristics for each of the four levels for all six subscales. Table 2 presents an abbreviated version of the Analytic scale.

The second form of response criteria, the Impressionistic rating scale, required raters to make a judgment regarding the overall quality of each writing sample as an example of effective exposition. The rating rubric directed raters to assign each essay a single numerical score by employing a six-point scale, with six (high) and one (low). In addition to the scale, the rubric contained several definitions outlining prominent conventionally recognized features of exposition.

Additionally, an indirect measure, a 37-item multiple-choice test, was developed to measure skills presumably related to actual production of expository writing. The skills covered in the test were identified through two related analyses. The first analysis consisted of a review of expository writing skills frequently emphasized in secondary composition curricula and instructional materials for which selected-response type practice was provided. For example, a typical exercise required students to identify from a list those details which either do or do not support a given generalization. The skills identified through the review were then arrayed against elements listed in the Analytic rating scale in order to determine which skills were conceptually analogous to the rating scale elements.

The final version of the Objective measure contained a subtest for each of the following Analytic scale elements: focus (main idea), development, organization, support, paragraphing. Each of the five subtests contained five similar-format items which were generated according to a set of test item specifications. Objectives for the subtests are presented in Table 3. The method by which items for a sixth subtest were generated was different from that of the first five subtests.

Unlike the stimulus passages of the first five subtests, which were generated specifically for the objective test, the passages in subtest six were drawn from actual samples of students' writing. The passages selected exhibited one or more of the several types of errors, e.g., failure to state or imply the main idea, lack of supporting statements. Each passage was followed by four items directing the student to identify the statement(s) which exhibited a specified category of error.

Summary of Findings

Relationships Among Analytic Scale and Impressionistic Scale Scores

Correlations among scores from the two rating scales are presented in Table 4. As shown, scores on the six subscales comprising the Analytic scale proved to be highly related, with correlations ranging from .69 to .90. Correlations between subscale scores and the Analytic scale total scores ranged from .82 to .96.

The correlation between the Impressionistic scale scores and Analytic total scale scores (.81) indicated a strong association between the two rating strategies. The association extended to relationships between the six Analytic subscales and Impressionistic ratings, with correlations ranging from .65 to .80.

To further examine the relationship between Impressionistic and Analytic scores, Impressionistic scores were regressed on the six Analytic subscales, which jointly accounted for approximately 75% of the variation ($F = 53.736$, df = 6,105, p .01) in Impressionistic scores. Two Analytic subscales, Mechanics ($F = 30.789$) and Support ($F = 18.365$), proved to be significant predictors to Impressionistic scale scores in the model with all six subscales (see Attachment 7).

The relatively strong associations among the Analytic subscales suggested that the relative importance of the subscales as predictors may be masked in the regression analyses. To examine this, a new composite variable, Structure, which was comprised of the sum of scores on four Analytic subscales (Organization, Focus, Paragraphing, Development) was entered into the equation. The combination of the three subscales Mechanics, Support, and Structure accounted for 74% of the variation in Impressionistic scale subscores ($F = 105.159$, $df = 3,108$, p $.01$). Again, Support ($F = 28.502$) and Mechanics ($F = 31.468$) emerged as significant predictors to Impressionistic scores (see Table 5).

Relationships Among Rating Scale Scores and the Objective Measure

The correlation between the Objective measure total scores and the Analytic scale total scores was .61, while the correlation with Impressionistic scores was .65 (see Table 6). Correlations between the six Objective subtest scores and total scores on the two rating scales ranged from .55 to .23.

To examine the predictive relationship between the Objective measure and the two rating scales, the Impressionistic and Analytic scale total scores were independently regressed on the six major Objective subtests (see Table 7). Results of the regression analyses for the Impressionistic scale ($F = 12.853$, $df = 6,92$, p $.01$) and for the Analytic scale ($F = 10.338$, $df = 6,92$, p $.01$) indicated that the six Objective subtests jointly accounted for approximately 46% of the variation in Impressionistic scale scores and for approximately 40% of the variation in Analytic scale scores. Two Objective subtests, Paragraphing (Subtest 4) and Paragraph Analysis (Subtest 6), proved to be significant predictors to both Impressionistic and Analytic scale total scores.

In an additional series of analyses, Analytic subscales were regressed on Analogous groups of Objective items. These calculations (reported in Table 8) resulted in significant F ratios for eight of the nine groups of Objective items.

Discussion

An interesting finding to emerge from the study concerned the pattern of strong relationships across the Analytic subscales. On the basis of these data alone, it is tempting to infer that the Analytic scale actually tapped a single unitary dimension of writing. Such an inference, however, overlooks an important facet of the writing task which may have affected the results--the writing topic and directions.

The topics and directions of the study were specifically designed to elicit writing developed through a logically arranged structure of generalizations supported by specifics. Directions accompanying the topics prompted students in the following way:

> Remember, the purpose of your essay is to give an informative explanation...Back up your ideas with specific support, such as examples, facts, and other details. Make sure your essay is well-organized.

As anticipated, the topics and directions promoted uniformity in the rhetorical structures of the majority of students' writing samples, of course, with varying degrees in quality of execution. Whether differential subscale scores will emerge when writing samples display more varied structural patterns is an area worthy of additional inquiry.

It is also possible that Impressionistic rating scores were indirectly a result of the relative uniformity of the structural characteristics of the essays. Once Impressionistic raters became habituated to the structural patterns of the majority of writing samples (through practice and training), they may have attended to a few prominent features--other than structure--which most noticeably discriminated among the essays.

The preceding notion is supported by the emergence of the two Analytic subscales, Mechanics and Support, as predictors to Impressionistic ratings. Judgments regarding a writer's command over mechanical aspects of writing can be made independently of the overall structure of a written product. Similarly, judgments regarding adequacy of support for generalizations embedded within an essay can be made without reference to the overall structure of a written piece. Given the relative conformity of structural patterns of the students' essays, Analytic and Impressionistic raters may have inadvertently attended to two readily discernible, and thus discriminating, features of the writing samples: mechanics and support.

This interpretation of statistical relations between the two rating strategies is not meant to imply that the discourse domain of exposition consists exclusively of two components, mechanics and support. The high correlations between Analytic scores and Impressionistic scores suggest that the Analytic subscales did, in fact, represent recognizable, if not necessarily independent, features of exposition. A provocative issue presents itself: Was the relative lack of independence among the Analytic subscales--and the high correlation with Impressionistic scores--a function of the homogeneity of student responses? Or, are features of writing such as development, organization, and main idea actually inseparable for purposes of rating?

Not too surprising were findings which revealed moderately positive relationships between Objective measure scores and those yielded by the two essay rating strategies. The positive relationship was expected, as the Objective items were designed to assess, at the levels of recognition and discrimination, those categories of skill measured by the rating strategies at the level of production. Moreover, previous studies have demonstrated that reasonably well-designed objective tests of writing invariably correlate with scores on writing samples, a phenomenon which, commonsensically, can be accounted for by the global constructs of language (reading) ability or verbal aptitude.

Of more compelling interest to competency-based test developers were the patterns of student performance on the Objective measure. Given that reading ability was likely to affect test performance, the majority of items were developed with the aim of minimizing reading difficulty, e.g., avoiding complex constructions, abstract content, advanced vocabulary. In fact, most of the items were designed to require students to make discriminations among individual sentences, rather than sentences embedded within prose passages.

The relatively high mean performance (see Table 9) of students on items requiring discrimination among individual sentences suggested that many of the students possessed the writing-related competencies being measured, such as selecting details to support generalizations, arranging given ideas in a logical order, choosing statements to develop a given main idea. For many lower-ability students (as indicated by teacher ratings), though, these competencies were not expressed when stimulus competition within the task (e.g., number of words and sentences, sentence structure) was increased.

Especially worthy of consideration are properties of items within the two subtests (Subtests 4 and 6) which proved to be significant predictors to the essay-rating total scores. Both of these subtests required students to make discriminations among statements embedded within prose passages. Here again, the construct of verbal (or reading) ability is a convenient, but hardly satisfying, explanation for a statistical artifact.

The preceding discussion has highlighted some of the technical problems and issues associated with the measurement of writing and writing-related skills, but what about practical implications? Findings indicated that the two rating scales provided essentially comparable estimates of writing competence. By definition, analytic scales, such as the one employed, have a greater potential than impressionistic (or holistic) scales to paint a clear picture of students' writing strengths and deficits. As expected, the time required to train Analytic raters and to analytically rate the essays was slightly greater than the time required for Impressionistic procedures (approximately six hours). The expense, however, is likely to be outweighed by the usefulness of the information yielded by analytic procedures. The explicit nature of analytic scales provides instructional decision makers and students with clear information on the domain of learning being measured. Such information is likely to enhance dialogue among teachers, students, and administrators on the status of student writing and may serve as a basis for instructional planning, diagnosis, and remediation.

The contribution of indirect measures of writing, however, raises a variety of issues regarding further study. One of the most basic issues centers on the nature of the relationship between direct and indirect measurement strategies. As demonstrated in the present study, as well as others, there exists an array of selected response tasks (beyond those measuring sentence level skills and writing mechanics) which are conceptually and statistically associated with writing production. If selected response tasks are to be useful within the context of competency-based measurement, though, test developers must employ test items which are positively related to instructional efforts.

Eighteen years ago, Richard Braddock characterized research on composition as ". . . laced with dreams, prejudices, and makeshift operations" (Braddock et al., 1963). It probably is fair to say that the state of composition research has advanced, even accelerated, in recent years. This may be due in large part to growing acceptance of a research paradim which views measurement and instruction as complementary pursuits. A continuing challenge to test developers, then, is to identify measurement tasks which, when practiced under appropriate instructional conditions, are likely to promote, not simply predict, acquisition of writing production skills.

REFERENCES

Braddock, R., Lloyd-Jones, R., & Schoer, L. Research in written composition. Champaign, Ill.: National Council of Teachers of English, 1963.

Brooks, C., & Warren, R. Modern rhetoric, shorter edition. New York: Harcourt, Brace & World, 1961.

Godshalk, F. I., Swineford, F., & Coffman, W. E. The measurement of writing ability. New York: College Entrance Examination Board, 1966.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, Calif.: McCutchan, 1974.

Office of the Los Angeles County Superintendent of Schools. A common ground for assessing competencies in written expression, review copy. Los Angeles, Division of Curriculum and Instructional Services, 1977.

TABLE 1

Measurement Strategies

Type of Strategy:	
Direct Measure of Writing	Indirect Measure of Writing
1. Writing samples judged by Analytic scoring criteria.	3. Objective items requiring students to discriminate among given passages of written discourse.
2. Writing samples judged by Impressionistic scoring criteria.	

TABLE 2

Analytic Rating Scale (Abbreviated)

THE ANALYTIC RATING SCALE

Analytic Scale Elements

1. Essay Focus: The introduction or conclusion of the essay clearly indicates the subject and main idea of the whole essay.
2. Essay Development: All major subtopics ("main points") clearly relate to the main idea of the whole essay.
3. Essay Organization: The main idea is developed according to a clearly discernible method of organization.
4. Support: Generalizations and assertions are supported by specific clear supporting statements.
5. Paragraphing: The essay is composed of one or more clearly discernible units of thought, e.g., paragraphs.
6. Mechanics: The essay is free of intrusive mechanical errors.

TABLE 2 (con't.)

Analytic Rating Scale: Sample Rubric

ELEMENT 1

Essay Focus

The introduction (if deductively structured) or conclusion (if inductively structured) of the essay clearly indicates the subject and main idea of the whole essay.

4. The introduction (and/or conclusion) of this paper clearly conveys the main idea of the whole essay. It also limits the topic by alerting the reader to the key points covered in the body of the essay.

   Specifically, in the introduction (and/or conclusion):

   a. The subject of the essay is clearly identified.

   b. The main idea of the whole essay is clearly stated or implied.

   c. The topic is clearly limited. That is, key points (e.g., reasons, ideas) or major line(s) of reasoning treated in the essay are identified or summarized.

3. The introduction (and/or conclusion) of this paper conveys the main idea of the whole essay. It sets limits on the topic, but does not clearly suggest how the main idea is developed.

   Specifically, in the introduction (and/or conclusion):

   a. The subject of the essay is clearly identified.

   b. The main idea of the whole essay is clearly stated or implied.

   c. An attempt is made to limit the topic. That is, the number -- or type -- of key points is specified, but there is not clear reference to the substantive issues treated in the body of the essay.

2. The introduction (and/or conclusion) of this paper gives the reader a fairly clear sense of the main idea of the whole essay. However, neither the introduction nor the conclusion help focus -- or bring direction to -- the body of the paper.

   Specifically, in the introduction (and/or conclusion):

   a. The subject of the essay is identified.

   b. The main idea of the whole essay is stated or implied.

   c. No attempt is made to limit the topic.

TABLE 2 (con't.)

ELEMENT 1 (continued)

Essay Focus

1. Neither the introduction nor the conclusion is helpful to the reader in obtaining any sense of the main idea of the essay.

   Specifically, in the introduction (and/or conclusion):

   a. The subject of the essay is not clearly identified or there is no reference to the subject

   AND/OR

   b. The main idea of the whole essay is not clearly stated or implied or no reference is made to the main idea, or the reference is confusing.

TABLE 3

Objective Measure Subtest Objectives (Abbreviated)

I. Objective

(5 items) The student will be given a brief paragraph which is lacking either a topic or concluding sentence, and four alternate sentences. The student is to select the sentence which would serve as the most appropriate topic or concluding sentence, i.e., a statement of the paragraph's main idea.

II. Objective

(5 items) The student will be given a main idea statement for a multi-paragraph expository essay and alternate statements that might be included in the body of the essay. The student is to select the statement that most directly contributes to development of the given main idea.

III. Objective

(5 items) The student will be given a topic sentence for an expository paragraph and alternate statements that might be included in the paragraph as supporting detail. The student is to select the statement that does not provide specific support for the given topic sentence.

IV. Objective

(5 items) The student will be given a series of five to six lettered sentences which express two distinct thoughts (i.e., sub topics) which are related to the same overall main idea. The student is to indicate where one complete thought ends and another begins, i.e., where a new paragraph could logically begin.

V. Objective

(5 items) The student will be given five sentences which could be included in an expository essay: One statement of the essay's main idea; two "sub topic" sentences ("topic" or "concluding" sentences for individual paragraphs within the essay); two supporting details. The sentences will be given in scrambled order. The student is to indicate a logical order for the sentences.

TABLE 4

Regression of Impressionistic Scale on Analytic Scale

REGRESSION OF IMPRESSIONISTIC SCALE SCORES ON SIX ANALYTIC SUBSCALES

Analytic Subscales	Unstandardized Coefficient	Standard Error of β	Standardized Coefficient	F
Mechanics	.585	.106	.424	30.789*
Focus	.233	.147	.196	2.537
Development	.046	.160	.040	.082
Organization	.293	.182	.265	2.492
Support	.514	.120	.424	18.365*
Paragraphing	.033	.148	.028	.048

df = 1,105

* = $p < .01$

$r^2 = .75$

R = .87

TABLE 5

REGRESSION OF IMPRESSIONISTIC SCALE SCORES ON THREE ANALYTIC SUBSCALES

Analytic Subscales	Unstandardized Coefficient	Standard Error of β	Standardized Coefficient	F
Mechanics	.582	.104	.422	31.468*
Support	.560	.105	.462	28.502*
Structure[1]	.017	.028	.056	.370

df = 1,108

* = $p < .01$

Structure[1] = Composite score of Organization, Focus, Development, Paragraphing

$r^2 = .74$

R = .86

TABLE 6

CORRELATIONS AMONG ANALYTIC AND IMPRESSIONISTIC TOTAL SCORES AND OBJECTIVE MEASURE SCORES

			I	II	III	IV	V	VI	A	B	C	D	
	Imp. Total	An. Total	Main Idea	Development	Supt. Detail	Paragraph	Organization	Para. Analysis	Main Idea[1]	Development[1]	Organization[1]	Supt. Detail[1]	Total Test
Impressionistic Total	1.00												
Analytic Total	.81	1.00											
Objective Subtests													
I. Main Idea	.24	.23	1.00										
II. Development	.46	.37	.17	1.00									
III. Supt. Detail	.35	.35	.11	.35	1.00								
IV. Paragraph	.55	.49	.29	.43	.37	1.00							
V. Organization	.45	.42	.26	.38	.32	.39	1.00						
VI. Para. Analysis	.53	.53	.36	.47	.35	.43	.43	1.00					
A. Main Idea[1]	.30	.35	.24	.36	.12	.23	.12	.72	1.00				
B. Development[1]	.40	.40	.26	.33	.30	.35	.37	.74	.40	1.00			
C. Organization[1]	.39	.33	.32	.35	.35	.32	.33	.75	.44	.34	1.00		
D. Supt. Detail[1]	.47	.51	.21	.33	.24	.33	.42	.71	.30	.38	.44	1.00	
Objective Total Test	.65	.61	.53	.66	.55	.68	.68	.85	.54	.66	.67	.63	1.00

1 = Items within Paragraph Analysis

TABLE 7

Regression of Rating Scales on Objective Measure

REGRESSION OF ANALYTIC SCALE TOTAL ON SIX OBJECTIVE MEASURE SUBTESTS

Objective Measure Subtests	Unstandardized Coefficient	Standard Error of β	Standardized Coefficient	F
Main Idea	.141	1.300	.009	.012
Development	.428	1.595	.026	.072
Supporting Detail	1.502	1.669	.082	.811
Paragraphing	4.037	1.514	.260	7.105*
Organization	1.771	1.173	.142	2.277
Paragraph Analysis	2.184	.674	.330	10.488*

df = 1, 92
* = $p < .01$
R = .66
r^2 = .40

REGRESSION OF IMPRESSIONISTIC SCALE TOTAL ON SIX OBJECTIVE MEASURE SUBTESTS

Objective Measure Subtests	Unstandardized Coefficient	Standard Error of β	Standardized Coefficient	F
Main Idea	.025	.274	.007	.008
Development	.497	.336	.137	2.182
Supporting Detail	.209	.352	.051	.354
Paragraphing	1.06	.319	.310	11.088*
Organization	.421	.248	.153	2.894
Paragraph Analysis	.371	.142	.254	6.08*

df = 1, 92
* = $p < .01$
R = .68
r^2 = .46

TABLE 8

Regression of Analytic Rating Subscales on Groups of Objective Items

REGRESSION OF ANALYTIC SUBSCALES ON ANALOGOUS OBJECTIVE MEASURE SUBTESTS

Analytic Subscale	Objective Subtest	Unstandardized Coefficients	Standard Error of β	Standardized Coefficients	F	df	R	r^2
I. Focus	Main Idea	.518	.269	.188	3.709*	1,96	.37	.14
	Main Idea[1]	1.051	.375	.273	7.826*	1,96		
II. Development	Development	.660	.307	.213	4.619*	1,96	.40	.17
	Development[1]	.953	.333	.283	8.171*	1,96		
III. Support	Supporting Detail	.545	.313	.161	3.026*	1,96	.48	.23
	Supporting Detail[1]	1.548	.347	.412	19.896*			
IV. Paragraphing	Paragraphing	1.303	.269	.442	23.511*	1,97	.44	.20
V. Organization	Organization	.905	.360	.360	13.516*	1,96	.42	.18
	Organization[1]	.514	.131	.131	1.797	1,96		

1 = items within Subtest VI, Paragraph Analysis

* = $p < .01$

ALTERNATIVE SCORING SYSTEMS FOR PREDICTING CRITERION GROUP MEMBERSHIP

Lynn Winters

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California 90024

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

ABSTRACT

The purpose of this study was to describe the effects of using competing scoring systems on classification decisions involving high school and college writers. Two questions were of interest:

1. What are the comparative reliabilities of different scoring systems?
2. What are the comparative validities of different systems for classifying writers into the appropriate levels of writing skill?

THE EFFECTS OF DIFFERING RESPONSE CRITERIA ON THE ASSESSMENT OF WRITING COMPETENCE

The method selected for judging writing samples *ipso facto* defines "good writing." However, decisions about writing competence derived from one scoring system may not lead to the same decisions about an examinee when another scoring system is used. The effect of using alternative methods of judging compositions on the classification of examinees into *a priori* defined levels of writing skill has not been investigated. The purpose of this study was to describe the effects of using competing scoring systems on classification decisions involving high school and college writers. Two questions were of interest:

1. What are the comparative reliabilities of different scoring systems?
2. What are the comparative validities of different systems for classifying writers into the appropriate levels of writing skill?

The Validation of Scoring Systems

Types of Scoring Systems

Scoring systems are generally of three types: impressionistic, analytic, or frequency counts (Braddock et al., 1963). General impression marking is the system whereby two or more raters quickly read an essay then assign a single score ranking the writing sample in relation to all other papers being evaluated. The several ratings assigned by different scorers are averaged to obtain a reliable (i.e., stable) estimate of a writer's ability. Several sorts of general impression marking are popular and incorporate rubrics of various levels of specificity and different score ranges. The one feature shared by impressionistic systems is the inclusion of rater training prior to scoring in order to "calibrate" readers in the assignments of marks (Conlan, 1976).

Analytic scoring systems, in contrast to general impression marking, require careful scrutiny of the writing samples. While impressionistic systems are predicated on the assumption that good writing is clearly recognizable, analytic systems assume that quality writing is characterized by the inclusion of certain rhetorical elements. One of the most popular of the analytic rubrics is the ETS Composition Scale, often called the Diederich Expository Scale. This scale emerged from a study that investigated factors influencing human judgment of expository writing samples. A factor analysis of patterns of rater agreement, combined with a review of rater comments on the essays belonging to each of five obtained clusters, was used to form the categories for the Diederich Expository Scale (DES).

The third type of scoring system is the frequency count. Whereas general impression and analytic scales focus on the quality of an essay, frequency counts are concerned with the quantity of such linguistic elements as the number of unique words in an essay, the number of sentences, or the number of words per independent clause. The frequency counting method with perhaps the most potential for judging levels of writing skill is the T-unit. T-units are the shortest grammatically complete sentences that a passage can be cut into without creating fragments (Hunt, 1977). Studies investigating the relationship of T-unit length to some measure of writing quality have tended to support the hypothesis that a positive relationship between the two exists (Howerton, 1977; Hunt, 1977; O'Hare, 1973).

A number of studies (Follman & Anderson, 1967; O'Hare, 1973; Howerton, Jacobson, & Elden, 1977) have found relationships among the three types of scoring systems. These findings have led researchers to hypothesize that different scoring systems measure a number of elements in common (Follman & Anderson, 1967). However, a simultaneous comparison of impressionistic, analytic, and frequency count systems has not been done. The different assumptions about writing skill implicit in each type of rubric suggest that high intercorrelations may not exist. If such is the case, their interchangeability for judging writing competence is impugned.

Validation Procedures

There have been few attempts to validate direct measures of writing skill (i.e., sets of response criteria) beyond the level of content or descriptive validity. The content validity of the scoring criteria used in NAEP and the College Entrance Examination Board's Advanced Placement Examination was established by "expert judgment" (Lloyd-Jones, 1977). Neither of these well-established systems has been investigated as to its ability to predict success on subsequent tasks or classify examinees, issues in empirical validation.

While few guidelines exist in the literature on writing research for the empirical validation of scoring systems, techniques for establishing empirical validity have been described in the context of criterion-referenced testing (Hambleton et al., 1977; Harris, 1974; Millman, 1974). Criterion-referenced measurement technology offers two major strategies for investigating empirical validity: (1) the use of scores on established tests as a criterion to which newly developed test items should predict, and (2) the use of the performance data of criterion groups to validate the accuracy of classification decisions to be made from the test.

The use of criterion groups to investigate the empirical validities of different scoring criteria provides an *a priori* classification system against which the criteria can be tested, i.e., the ability to sort examinees representing different levels of writing skill into the correct writing groups. The criterion groups in this study were chosen to contain low and high performance high school and college writers. These four groups provided an operational definition of the construct "writing competence." The four scoring systems validated were selected to represent the three types of scoring criteria: impressionistic, analytic, and frequency counting. Two types of analytic scales, the Diederich Expository Scale and the CSE Analytic Scale, were tested. The Diederich Scale contains both impressionistic elements and elements tied to the requirements of exposition. The CSE Analytic Scale was derived from current composition textbooks and emphasizes the elements of writing taught in the schools. T-unit analysis was the method selected to exemplify the frequency counting approach. It was chosen because it was validated against qualitative measures of writing and was expected to have elements in common with impressionistic and analytic scales.

Procedures

Assignment of Papers to Raters

Two designs were used for assigning student papers to raters, one for scoring sample papers used to evaluate the effectiveness of rater training (Pilot Studies) and the second to examine the writing performance of the four criterion groups across the four scoring systems (Final Papers). Eighty subjects produced two expository essays on two test occasions, one week apart. Two sets of four raters were trained in two scoring systems and graded essays after each training session.

The Pilot Studies were used to insure that all raters were assigning essay scores consistent with each other, to investigate topic effects, and to obtain estimates of generalizability for several research designs for the final rating of papers. These studies were conducted four times, once after each rater training session. Twenty-four papers, chosen to represent all four criterion groups and both topics, were scored by all four raters. Two analyses of variance were run on the pilot scores, one to obtain intraclass coefficient alpha and the other to obtain estimates of variance components for subjects, topics, raters, and their interactions under a random model, fully crossed design. The coefficients calculated from these analyses are displayed in Table 1 below.

Table 1

Alpha and Generalizability Coefficients: Pilot Data

Coefficient and Condition	Scoring System			
	GI	DES	CSE	T
Alpha: 4 Raters--Topics Fixed	.97	.82	.96	.98
Generalizability: 4 Raters, 2 Topics	.87	.88	.80	.65
Predicted G: 2 Raters, 2 Topics	.85	.79	.78	.47
Predicted G: 2 Raters, 1 Topic	.85	.87	.85	.68

Results of the pilot studies indicated that fairly high reliability could be obtained under a nested design with two raters reading one paper per examinee. The final rating design crossed subjects and topics with raters, but nested raters within topics. Each rater read only 80 papers, one from each subject, 40 of which were Topic 1, and 40 of Topic 2. Subjects and topics were randomly assigned to raters.

Figure 1

Assignment of Writing Samples

	Group			
Rater	HSL n=20	HSH n=20	Coll L n=20	Coll H n=20
1	T_1 n=10	T_1 n=10	T_1 n=10	T_1 n=10
	T_2 n=10	T_2 n=10	T_2 n=10	T_2 n=10
2				
.				
.				
6				

Subjects

The subjects of the study were high school and college students enrolled in composition classes in June 1981. Twenty (20) sets of writing samples were randomly selected from papers obtained from writing classes at a suburban, working-class high school and a major urban university. Classes were chosen to represent two levels of writing skill, high and low, for both high school and college populations. The students whose samples were selected became the criterion groups described below:

1. Low Performance High School Writers (HSL)

   Students enrolled in basic composition classes at the eleventh grade level who do not intend to attend a university.

2. High Performance High School Writers (HSH)

   Students enrolled in eleventh grade advanced composition classes who plan to attend a university.

3. Low Performance College Writers (Coll L)

   Students enrolled in remedial college composition at a university by virtue of having failed a written English placement examination.

4. High Performance College Writers (Coll H)

   Students enrolled in freshman composition at a university after having passed the written placement examination or a course in remedial composition.

Writing Tasks

Subjects were given 50 minutes to produce a 200-word expository sample on two test occasions, one week apart. Two parallel topics were randomly assigned to subjects and counterbalanced to control for test occasion. Dictionaries were not available, although subjects were told to revise papers if time allowed.

Controls

The design did not attempt to control for the instructional history of the subjects or the effects of test occasion. It did, however, attempt to control for the following: (1) topic effects, (2) rater background, and (3) inter-rater disagreement. Table 2 and Figure 2 summarize the steps taken to control sources of variation due to topics and raters.

Dependent Measures and Rater Training

There were four dependent measures for each subject. These consisted of each student's score over two topics and two raters for each of the four scoring systems. As the validity of these measures was

Table 2

A Summary of Controls for Sources of Variation Due to Topic and Rater

1. Examinee scores were averaged across topics.
2. A stratified random sampling procedure was used so that each rater received one writing sample from each of the 80 subjects and an equal number of papers on each topic.
3. Papers were arranged in rater packages so that each rater received papers in random order. Additionally, rater packets were created so that there was no systematic pairing of raters.
4. Writing samples were typed and assigned random identification numbers to obscure examinee name and criterion group status.
5. To minimize reactivity, scoring systems were assigned to training conditions so that the two analytic scales were separated and general impression marking introduced first.
6. Each training condition consisted of four raters, two high school and two college teachers. All raters learned at least two systems. Two raters were trained in all four systems.

Figure 2

Assignment of Raters to Training Conditions

	Session 1 GI & DES		Session 2 CSE & T
Rater 1:	College remedial teacher	Rater 1:	College remedial teacher
Rater 2:	College composition	Rater 2:	College composition
Rater 3:	High school English	Rater 3:	High school English
Rater 4:	High school English	Rater 4:	High school English

directly dependent upon their reliabilities, raters were trained to use each system before being allowed to score essays. The general procedure for rater training in each system follows: (1) Raters were given materials explaining the purpose of the scoring system, copies of the topics, and practice papers representing samples of each of the criterion groups. (2) Raters discussed the rubric and defined scoring categories among themselves. (3) Practice papers were scored and discussed until raters felt they could agree on the assignment of marks. (4) Raters read 24 sample papers for a pilot study (described in a previous section). (5) When the alpha coefficient from the pilot data was .80 or better, raters were deemed "trained" and allowed to score the actual writing samples.

In investigating the classification validities of the scoring sytems, two issues were involved: Which systems discriminated the best among groups? And which systems classified subjects most accurately? For each of these issues, the dependent measures differed. The following outline summarizes the variables of interest for each question:

I. Variables of Interest in Finding the System(s) Which Best Discriminate Among Groups

    A. Independent Variables: Criterion Groups

    B. Dependent Variables: Scores Obtained on Systems

        1. GI: Score Scale 1-6
        2. DES: Score Scale 5-45
        3. CSE: Score Scale 6-24
        4. T: Score Scale 8-34 words per T-unit

II. Variables of Interest in Finding Scoring Systems Which Best Predict Criterion Group Status

    A. Independent Variables: Scoring Systems as a Treatment

    B. Dependent Variables: Criterion Group Membership

Analyses

Several preliminary analyses were necessary in order to select the appropriate procedures for analyzing data related to the question of classification validity. Rater agreement and generalizability coefficients were calculated. The relationship among scoring systems was examined with Pearson Correlations. Finally, univariate analyses of variance for each scoring system were run to identify those systems which distinguished among writing groups (see Tables 3-6).

The results of these analyses indicated that while all systems had high alpha coefficients, the reliability coefficients were somewhat lower and not as comparable. T-unit, which had the highest alpha (.99) could not be investigated with a G coefficient due to unusual subject by topic interactions. DES had the highest generalizability coefficient (.85) and an alpha of .90. GI and CSE were comparably reliable with Gs of .63 and .67 respectively and alphas of .81 and .97. Reliability dropped from the predicted pilot coefficients. Finally, unusual subject by topic effects of a high magnitude were found. These could not be explained because the research design confounded topic with occasion for the individual examinee and rater with topic.

Intercorrelations among the three systems for which significant differences among writing groups were found (GI, DES, CSE) at first appeared quite high. However, when the data were analyzed by criterion group, the relationship among these systems fluctuated from mild to weak. T-unit analysis did not measure what is conventionally described by holistic systems (GI) or analytic criteria (DES, CSE) as good writing. Its negative relationship to other systems for some of the criterion

Table 3

Data for Estimating Generalizability Coefficients: Final Papers

Source	SS	DF	MS	Estimated Variance Components*	
(a) General Impression					
Subjects	224.286	57	3.935	.418	
Topics	3.446	4	.862	0.000	G=.63
Raters within Topics	3.941	8	.492	.003	
Subjects X Topics	55.304	57	.970	.325	
Residual	36.556	224	.321	.321	
Subjects by Rates within Topics					
(b) Diederich Expository Scale					
Subjects	9649.122	52	185.560	33.473	
Topics	647.993	4	161.998	0.000	G=.85
Raters within Topics	76.227	8	9.528	.073	
Subjects X Topics	1202.453	52	23.124	8.852	
Residual	563.693	104	5.420	5.420	
Subjects by Rates within Topics					
(c) CSE Analytic Scale					
Subjects	2447.302	62	39.473	4.400	
Topics	30.800	4	7.7	0.000	
Raters within Topics	23.092	8	2.887	.025	G=.67
Subjects X Topics	545.944	62	8.806	3.773	
Residual	156.376	124	1.261	1.261	
Subjects by Rates within Topics					
(d) T-Unit Analysis					
Subjects	1370.150	51	26.866	0.000	
Topics	290.958	4	72.740	.002	
Raters within Topics	2.042	8	.255	.001	G=Undefined
Subjects X Topics	1488.059	51	29.178	14.493	
Residual	19.537	102	.192	.192	
Subjects by Rates within Topics					

* Negative variance components are reported as zero.

Table 4

Intercorrelations Among Scoring Systems for Total Sample (N=80)

System	GI	DES	CSE	T-Units
GI	-			
DES	.82*	-		
CSE	.79*	.86*	-	
T-Unit	.00	.06	.00	-

* Indicates significant correlations at p .05, df 18.

Intercorrelations Among Scoring Systems by Criterion Groups (n=20)

	Low			High		
	GI	DES	CSE	GI	DES	CSE
High School						
GI	.64*			.54*		
DES	.71*	.79*		.44*	.30	
CSE	.28	.33	.30	.00	.24	.21
College						
GI						
DES	.53*			.51*		
CSE	.11	.37		.70*	.60*	
T	-.07	-.39	-.52	-.31	-.23	-.52

* Indicates signficant correlation at p .05, df 18.

Table 5

Criterion Group Performance Across Scoring Systems: Means and Standard Deviations

System	HS Low n=20	HS High n=20	Coll Low n=20	Coll High n=20	Total n=80
GI	x = 1.97	3.21	4.01	3.33	3.02
(Total 6)	S.D.= .50	.65	.72	.74	1.00
DES	x =13.25	25.66	28.00	26.00	23.29
(Total 45)	S.D.= 2.95	2.91	3.26	4.09	6.63
CSE	x = 9.84	15.50	17.07	16.18	14.65
(Total 24)	S.D.= 1.57	1.92	2.08	2.83	3.55
T-unit	x =14.18	15.07	13.99	14.89	14.52
(Range 10-34)	S.D.= 3.45	5.18	1.95	2.05	3.38

Table 6

Analyses of Variance for Each Scoring System: Criterion Groups as Independent Variables

System	Source	SS	DF	MS	F	Eta^2
GI	Between	752.637	3	250.879	36.094	.59
	Within	528.250	76	6.951		
DES	Between	42930.338	3	14310.113	80.522	.76
	Within	13506.550	76	177.218		
CSE	Between	10280.700	3	3426.090	46.320	.02
	Within	5622.500	76	73.980		
T	Between	263.740	3	87.91	.470	.02
	Within	14205.085	76	186.90		

groups suggested that, in the present study, information gained from T-unit scores would be of little benefit in predicting expository writing performance as defined by criterion group membership. It was therefore not used in later analyses.

The classification validities of the systems were investigated by a simple ranking procedure and with discriminant analysis. The discriminant analysis (see Tables 7 and 8) revealed only one significant Linear Discriminant Function (LDF) for each of two combinations of scoring systems, GI/DES/CSE and DES/CSE. An examination of the within-group correlation matrices and the standardized LDFs indicated that DES possibly contributed most to group separation when it was combined with either CSE or GI. Both DES and CSE separated group centroids better than GI.

Table 7

Within-Groups Correlation Matrix for Discriminant Analysis

System	GI	DES	CSE
GI	--		
DES	.54	--	
CSE	.48	.55	--
T	.01	.04	-.04

Table 8

LDF Coefficients for Two Discriminant Analyses:
GI/DE/CSE and DES/CSE

Analysis	Eigen Value	Relative % of Information Accounted for by LDF	Wilkes' Lambda	Standardized LDF Coeff.
GI/DES/CSE	3.388	96.69	.204	GI-0.61
DES/CSE	3.374	99.69	.226	DES-0.758 CSE-0.214 DES-0.793 CSE-0.232

Discriminant classification revealed a 61% accuracy of group placement when three systems were used (GI/DES/CSE) and a 59% accuracy with the two analytic systems. Classification accuracy was greatest for the High School Low group (95%) and least for the College High group (20%). Prediction was better for the Low writers in each level of school than the high. Misclassification tended to be from College High to College Low and College Low to High School High. The High School High group was misclassified as College High when only the two systems were used. When three systems were used, misclassifications were almost equally distributed between the two college groups for the High School High writers. The most accurate discriminant classification analysis is reproduced in Table 9 below:

Table 9

Classification Accuracy Using GI, DES, CSE scoring Systems

Actual Group	N	Predicted Group Membership			
		HSL	HSH	Coll L	Coll H
High School Low	20	95%	5%	0%	0%
High School High	20	0%	55%	10%	35%
College Low	20	0%	25%	70%	5%
College High	20	10%	20%	45%	25%

Percent of cases correctly classified: 61.25%

The frequency distributions used for examining group classification for each scoring system separately displayed the same trends as the discriminant classifications (see Figures 3-6 and Table 10). For all systems but T, the High School Low group was clearly discriminable from the other three groups. There was much intermingling among the other groups, although DES appeared to separate the High School High and College Low groups better than did the other two (GI/DES) effective

Table 10

Classification Accuracy for Scoring Systems

Actual Group	N	Predicted Group Membership			
		HSL	HSH	Coll L	Coll H
(a) General Impression					
High School Low	20	81%	19%	0	0
High School High	20	15%	36%	46%	0
College Low	20	0	21%	29%	56%
College High	20	5%	28%	39%	36%
Overall GI Accuracy = 46%					
(b) Diederich Expository Scale					
High School Low	20	93%	8%	0	0
High School High	20	2%	46%	38%	16%
College Low	20	0	17%	35%	48%
College High	20	7%	3%	49%	37%
Overall DES Accuracy = 53%					
(c) CSE Analytic Scale					
High School Low	20	88%	13%	0	0
High School High	20	3%	46%	36%	51%
College Low	20	0	23%	36%	51%
College High	20	10%	19%	39%	33%
Overall CSE Accuracy = 51%					
(d) T-Unit Analysis					
High School Low	20	38%	18%	15%	30%
High School High	20	30%	17%	28%	25%
College Low	20	25%	34%	16%	25%
College High	20	8%	32%	41%	20%
Overall T-unit Accuracy = 24%					

scoring systems. With all systems but T, it appears the High School Low subjects are assigned the bottom third of the scores with the other criterion groups almost normally distributed about the top two-thirds of the score rankings and virtually indistinguishable from each other.

Unexpected findings include the impotence of the T-unit to discriminate or classify, as well as its non-existent relationship to qualitative scoring systems. Also unexpected was the wide divergence in types of writing samples receiving high and low ratings from each system. A perusal of the actual essays, more than any other analysis, illustrated the differences among the four systems. There was an unanticipated reversal in group means for the College Low and High groups. Finally, the inability of the systems, acting singly or in concert, to clearly separate all four groups was unanticipated.

Conclusions

Classification Validity

In order for a scoring system to accurately classify examinees into the correct writing groups, it must be able to distinguish among the groups. It was found that three of the systems, General Impression, Diederich Expository Scale, and CSE Analytic Scale, were associated with differences in group performance. T-unit was not, making it virtually useless for classification purposes.

When the three "discriminating" scoring systems (GI, DES, CSE) were used to classify examinees into the proper criterion groups, it was found that three systems classified more accurately than two. This finding was predictable in that classification accuracy is often enhanced by the use of several sources of information. Another explanation for the advantage of using three scoring systems to classify examinees is tied to the kinds of scoring systems used in this study. Two analytic and a general impression system were retained for the discriminant analyses. Although one analytic system (DES) contributed most to group separation, the best prediction was obtained with a combination of DES and GI scores. It may be that accurate assessment of writing must incorporate both impressionistic and analytic strategies. Support for this interpretation is gained when it is remembered that the one system that appears as the strongest variable is the DES, a system combining both general impression and analytic elements. The CSE scale does not lead to as accurate classification of group membership as DES. This result may be due, in part, to the fact that the scale provides no rating category for an overall impression of the effectiveness of the essay.

These interpretations must be accepted with caution, however, due to the unusual intercorrelations of the systems. There are groups by scoring system differences which may diminish the possibilities that any one scoring system can be an accurate predictor for all four groups. It is quite apparent that either holistic or analytic methods are more effective for distinguishing among writers of several ability levels and within a restricted age range.

The results of the classification analyses indicate that it is fairly easy to identify the lowest ability writers in a group and even, perhaps, some of the highest, but most difficult to separate writers of medium to high ability.

High school grouping practices provide a reason for the better classification accuracies obtained for the high school than college groups. In high school, students are assigned to English classes on the basis of previous grades, teacher recommendation and post high school educational plans. The high school environment is smaller; teachers often know students well enough to "counsel" them into the appropriate English classes. On the other hand, the university composition placement decision is made on the basis of an impersonal, one-shot examination scored on criteria unknown to students and teachers. This procedure greatly increases the possibility for mistakes in classifying examinees.

Classification accuracy is best for the low writing groups in both high school and college. It may be that the characteristics of "low" writing performance are more stable and discernible than those of "high" writers. To paraphrase Tolstoy, poor essays are very much alike, but good essays are outstanding in their own ways.

Recommendations

Several limitations constrain the generalizability of this study's findings. Scores reported for each system were obtained by raters who had participated in training. Rater training causes persons with potentially divergent views to "agree to agree" on both the meaning of the scoring rubric and the meaning of the scores. Different raters could agree with each other but disagree with the interpretation of the scoring rubric and score assignments reported in this study. The norming phenomenon, though clearly unavoidable, is rarely reported in writing research. The lesson: scoring rubrics never specify the entire set of criteria used to judge an essay. There is always that "x" factor which stands for how a rater interprets the rubric.

A second limitation to the generalizability of results has to do with the subjects. Participants in this study do not represent a random sample of high school or college writers. They do represent a sample of typical writing groups about which decisions must be made for placement into writing programs.

Finally, the topics used in this study, which appear to have produced some unusual effects on writing performance, were expository. There is no evidence that other modes of discourse or even other genres of expository topics would have produced the same results.

In spite of these limitations, several recommendations, both methodological and substantive, emerged from the study. Several procedures used facilitated interpretation of results. The use of a generalizability study during pilot ratings made it possible to identify conditions responsible for inconsistencies in score assignment. The G study allowed various designs for the final study to be considered in terms of the trade-off between reliability and cost-effectiveness. The calculation of both rater and G coefficients clarified the amount of rater disagreement contributing to variation in examinee scores attributable to error. Rater training provided valuable insights as to how human factors interact with scoring criteria and essays to produce a final estimate of writing ability. Finally, a review of representative essays revealed how scoring systems differed in their definitions of "good writing."

Substantive results point to testable generalizations about the comparative validities of the four systems for describing and classifying the writing performance of high school and college students. General Impression scoring, a widely used screening procedure which is speedily trained and scored, may not be the best system for making placement decisions or for graduation competency. The lack of descriptive power of its rubric severely restricts the instructional utility of this system when compared to analytic scoring techniques. Results of this study indicated that, while not sufficient as a placement tool, GI scores may be a necessary part of any writing assessment procedure.

The Diederich Expository Scale, while requiring extensive training time, appears to be the best system for distinguishing between high and low performance writers at both the high school and college levels. Its mild correlations with GI and CSE analytic scores indicate that the DES may be providing two sorts of information, impressionistic as well as analytic. It was found that discrimination and classification accuracy were increased when the number of data sources were increased. These findings suggest that the complex construct "writing skill" might best be assessed by more than one approach.

The use of the CSE scale to distinguish among examinees at different writing skill levels is a highly reliable and promising technique. The fact that this scale contributes less to group discrimination than the DES is possibly explained by its lack of attention to any of the qualitative aspects of writing. A revision of the scale with a category for the holistic aspects measured by GI may increase the classification and descriptive powers of this system.

The ability of measures of syntactic maturity to assess writing quality is seriously questioned by the results of this study. The lack of correlation of T-unit scores with three qualitative systems indicates that claims for a relationship between writing quality and syntactic maturity, at least for writers within a restricted age range, must be seriously re-examined.

References

Braddock, R., Lloyd-Jones, R., & Schoer, I. Research in written composition. Urbana, Illinois: National Council of Teachers of English, 1963.

Conlan, G. How the essay in the CEEB English test is scored. Princeton, N.J.: Educational Testing Service, 1976.

Follman, J. C., & Anderson, J. A. An investigation of the reliability of five procedures for grading English themes. Research in the Teaching of English, 1967, 190-200.

Hambleton, R., Swaminathan, H., Algina, J., & Coulson, D. Criterion referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1977, 48(1).

Harris, C. W. Some technical characteristics of mastery tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced testing. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Howerton, M. C., Jacobson, M. & Eldon, R. The relationship between quantitative and qualitative measures of writing skill. Paper presented at the Annual Meeting of the American Educational Research Association, New York, April 1977.

Hunt, K. W. Early blooming and late blooming syntactic structures. In C. Cooper & L. Odell (Eds.), Evaluating writing. State University of New York at Buffalo, 1977.

Lloyd-Jones, R. Primary trait scoring. In C. Cooper & L. Odell (Eds.), Evaluating writing: Describing, measuring, judging. Urbana, Ill.: National Council of Teachers of English, 1977.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, Calif.: McCutchan, 1974.

O'Hare, F. Sentence combining: Improving student writing without formal grammar instruction. NCTE Research Report No. 15. Urbana, Ill.: National Council of Teachers of English, 1973.